TRANSLATION-MEDIATED STRESS RESPONSES: MINING OF RIBOSOME PROFILING DATA

Krzysztof Franaszek

This dissertation is submitted for the degree of Doctor of Philosophy at the University of Cambridge

Churchill College June 2017

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. The content of this dissertation has not previously been submitted for any degree at this or any other university.

Krzysztof Franaszek

Summary

Advances in next-generation sequencing platforms during the past decade have resulted in exponential increases in biological data generation. Besides applications in determining the sequences of genomes and other DNA elements, these platforms have allowed the characterization of cell-wide mRNA pools under different conditions and in different tissues. In 2009, Ingolia and colleagues developed an extension of high-throughput sequencing that provides a snapshot of all cellular mRNA fragments protected by translating ribosomes, named ribosome profiling. This approach allows detection of differential translation activity, annotation of novel protein coding sequences and variants, identification of ribosome pause sites and estimates of *de novo* protein synthesis. As with other sequencing based methodologies, a major challenge of ribosome profiling has been sorting, filtering and interpreting the gigabytes of data produced during the course of a typical experiment. In this thesis, I developed and applied computational pipelines to interrogate ribosome profiling data in relation to gene expression in several viruses and eukaryotic species, as well as to identify sites of ribosomal pausing and sites of non-canonical translation activity.

Specifically, I applied various control analyses for characterizing the quality of profiling data and developed scripts for visualizing genome-based (exon-by-exon) rather than transcript-based ribosome footprint alignments. I also examined the challenge of mapping footprints to repetitive sequences in the genome and propose ways to mitigate the associated problems. I performed differential expression analyses on data from coronavirus-infected murine cells, retrovirus-infected human cells and temperature-stressed *Arabidopsis thaliana* plants. Dissection of translational responses in *Arabidopsis thaliana* during heat shock or cold shock revealed several groups of genes that were highly upregulated within 10 minutes of temperature challenge. Analysis of the branches of the unfolded protein and integrated stress responses during coronavirus infection allowed for deconvolution of transcriptional and translational contributions. During the course of these analyses, I identified errors in a recently publicized algorithm for detection of differential translation, and wrote corrections that have now been pulled into the repository for this package. Comparison of the translational kinetics of the dengue virus infection in mosquito and human cell lines revealed host-specific sites of ribosome pausing and RNA accumulation. Analysis of HIV profiling data revealed footprint peaks which were in agreement with previously proposed models of peptide or RNA mediated ribosome stalling. I also developed a simulation to identify transcripts that are prone to generating RPFs with multiple alignments during the read mapping process. Together, the scripts and pipelines developed during the course of this work will serve to expedite future analyses of ribosome profiling data, and the results will inform future studies of several important pathogens and temperature stress in plants.

Table of Contents

List of Figures & Tables

Chapter One

Introduction

1 CHAPTER ONE: INTRODUCTION

1.1 Translation

Proteins comprise more than 50% of the dry weight of a cell and are major effectors of biochemical catalysis, metabolic regulation and structural organization. The production of new proteins is estimated to constitute half of the energy expenditure associated with rapid cellular growth (Ingolia, 2016). Proteostasis is mediated by the combined effects of several processes, including *de novo* synthesis, protein degradation, chemical modifications, sequestration, secretion and uptake (through exosomes, endocytosis and other mechanisms).

The cellular proteome undergoes constant renewal and, under situations of environmental stress, reorganization. Several studies have measured global protein turnover rates in yeast and mammals and reported the average protein half-life to be ~1.5 hr to 1–2 days, respectively (Belle et al., 2006; Price et al., 2010; Cambridge et al., 2011; Toyama et al., 2013).

The control of new protein synthesis represents a key focus for cellular regulation and amplification of cell signalling. Schwanhäusser et al. (2011) used metabolic pulse labelling to measure absolute mRNA and protein abundance and turnover for more than 5,000 mammalian genes. Their analysis found that ~40% of the variation in protein concentrations is due to mRNA abundances while ~54% is due to variability in translation efficiencies. The dynamic range of protein abundance in terms of individual protein species is from zero to several million molecules per cell (Ghaemmaghami et al., 2003; Picotti et al., 2009) while the range of observed mRNA molecules is much narrower – 1 to 89 molecules per cell (Holstege et al., 1998). Several groups have identified large variations in the translational efficiency of specific genes (Ingolia et al., 2009; Gerashchenko et al., 2012; Vogel and Marcotte, 2012; Gerashchenko and Gladyshev, 2014; Csárdi et al., 2015). Translation of pre-existing mRNAs changes protein levels more quickly than transcription of new mRNAs (in the timespan of several seconds versus minutes), and translation regulation has been found to play a prominent role in various stress responses (Spriggs et al., 2010; Gerashchenko et al., 2012; Liu et al., 2013; Shalgi et al., 2013; Reid et al., 2014; Andreev et al., 2015a; Sidrauski et al., 2015). Translational control can also limit protein synthesis to specific sub-cellular locations, such as the axons of neurons, where synaptic translation is required for long-term potentiation and memory formation (Buffington et al., 2014).

A thorough understanding of translation is important for synthetic and industrial biology applications as well as for biomedical reasons (Lin et al., 2016). Understanding of ribosome activity has been used to develop gene circuits that can produce proteins in controlled

proportions or under selective conditions (Ketteler, 2012). The modulation of translation levels in itself has already been demonstrated as a valid mechanism for therapeutic intervention (Zangi et al., 2013; Liang et al., 2016). Of the 7000+ rare genetic diseases classified by the National Organization for Rare Diseases, ~2400 distinct genetic disorders are caused by an allele containing a premature termination codon, and this information is being exploited for drug development (Karijolich and Yu, 2014). For example, nonsense mutation suppression by aminoglycosides is currently being investigated as an approach to treat Duchenne muscular dystrophy (Malik et al., 2010; Keeling et al., 2014). Aberrant activity of translation initiation factors can cause wide-spread changes that promote uncontrolled cell proliferation, angiogenesis and eventually metastasis (Chu et al., 2016). Cap-dependent translation initiation factor eIF4E has been extensively studied as a oncology drug target (Schwartz et al., 1990; Silvera et al., 2010; Jia et al., 2012; Fields et al., 2013; Andreev et al., 2015a). Regulation of eIF4E is thought to be the mechanism of control for many immune system genes, and this target has been studied in various pharmaceutical screening assays (Piccirillo et al., 2014; Hett et al., 2016). Macrophage activation by interferon-γ has been shown to be affected to a great degree by modulation of translation activity of various interferon-stimulated genes, a fact that has helped improved understanding of inflammation and potentially various auto-immune disorders (Su et al., 2015)

Understanding of the differences in bacterial and eukaryotic translation is of importance to antibiotic and antidote development. For example, the small molecules erythromycin and tryptophan inhibit bacterial pathogenesis by promoting translational stalling of bacterial ribosomes (Arenz et al., 2014; Bischoff et al., 2014; Johansson et al., 2014). Diptheria toxin (produced by *Corynebacterium diphtheria*) that can cause childhood death, blocks protein synthesis in eukaryotes by inhibiting ribosome translocation, a fact that can be used to screen antidote candidates during pre-clinical investigation (Berg et al., 2012).

Mutations in genes encoding protein components of the ribosome or factors in ribosome biogenesis have been shown to cause various diseases in humans, including Diamond-Blackfan anemia, 5q-syndrome, Shwachman-Diamond syndrome, Dyskeratos congenita, Cartilage-hair hypoplasia and Treacher-Colins syndrome (collectively referred to as ribosomapathies) (Nakhoul et al., 2014).

This text will review components and mechanisms of translation, discuss several recently-developed methodologies to interrogate translation, survey the computational ramifications of high-throughput studies of translation and preview several recent investigations into translational processes.

1.1.1 Components of the eukaryotic translation apparatus

A key element of proteostasis is the production of new proteins in order to replace degraded, malformed or secreted proteins, or in response to changes in the cellular environment. The uncatalysed rate of peptide polymerization is too slow for many of the needs of cellular metabolism, hence the functional need for the ribosome, which increases the rate of peptide bond formation by 10^6- to 10^7-fold (Beringer and Rodnina, 2007). The ribosome is the primary mediator of cellular protein synthesis and there are 10^6 to 10^7 of these macromolecules in a eukaryotic cell (Pechmann et al., 2013). The eukaryotic ribosome is a ~$4.2*10^6$ D nucleoprotein complex consisting of ~5500 nucleotides of RNA and 80 proteins (Dinman, 2009; Berg et al., 2012). The fully-formed 80S complex consists of a 60S large subunit and a 40S small subunit. The former contains 47 ribosomal proteins and the 3354 nucleotide (nt) 28S rRNA, 154 nt 5S rRNA and 120 nt 5.8S rRNA. Its main function is to catalyse peptide bond formation on a growing polypeptide chain (Yusupova and Yusupov, 2014). The small subunit consists of 32 proteins and the 1753 nt 18S rRNA, and it functions to decode messenger RNA (mRNA). Initial investigations into the structure of the 80S ribosomes were thwarted by its high-complexity and the difficulty in finding well-diffracting crystals for high-resolution X-ray crystallography studies (Yusupova and Yusupov, 2014). Up until 2010, information about the structural organization of 80S ribosomes was derived from fitting high-resolution crystallographic structures of 70S ribosomes onto intermediate resolution cryo-EM models (Yusupova and Yusupov, 2014). Since then, several groups have succeeded in crystalizing eukaryotic ribosomes under a variety of conditions (Ben-Shem et al., 2010, 2011; Klinge et al., 2011; Weisser et al., 2013; Khatter et al., 2015).

While a full description of the ribosome structure is outside the scope of this thesis, several key components merit description. This section will discuss key components of the translational apparatus (primarily the ribosome) as well as their functionalities, including the peptidyltransferase centre, the P, A and E sites, and the mRNA entry tunnel (

Figure 1.1). The peptidyltransferase centre (PTC) is located on the large subunit and catalyses the aminolysis of ester bonds to join amino acids as they are added to the nascent polypeptide chain and hydrolysis of the peptidyl-tRNA linkage during translation termination (Beringer and Rodnina, 2007; Simonović and Steitz, 2009). The resulting protein is threaded through the exit tunnel, which is thought to accommodate ~30 amino acids (or up to 60 in an

alpha helical conformation; Kramer et al., 2009). Biochemical data suggest that the exit tunnel topology

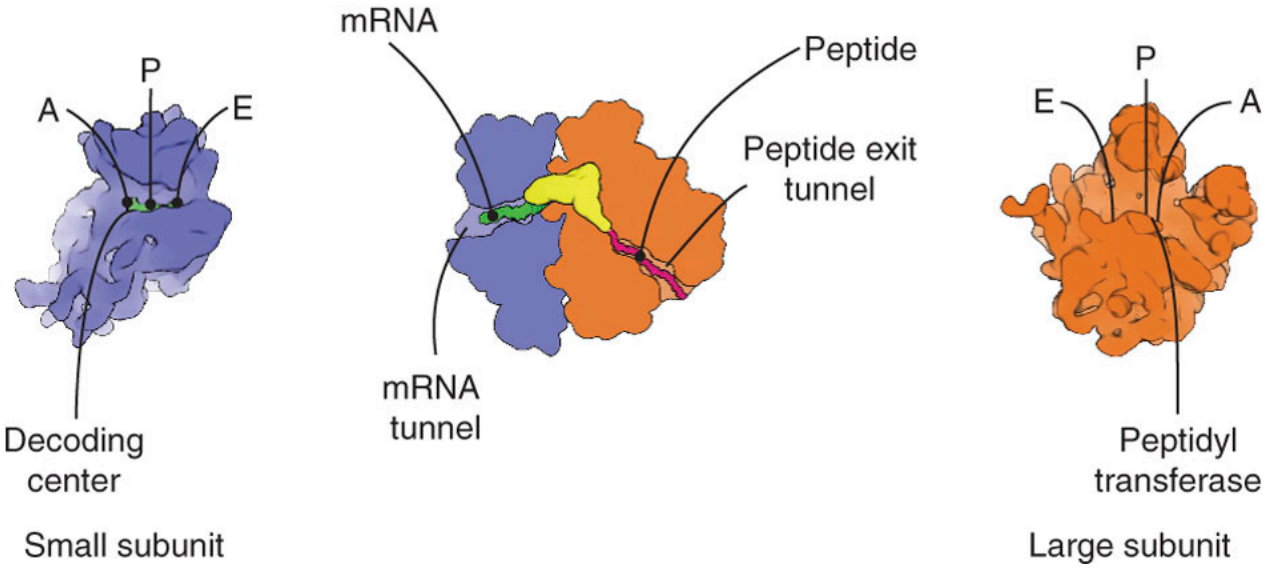


Figure 1.1 Key features of the ribosome. The left image indicates the small subunit and the right image shows the large subunit (both from the solvent side). The middle image shows the fully assembled ribosome loaded on an mRNA with a nascent peptide in the peptide exit tunnel. Figure adapted from Melnikov et al., (2012).

must be non-static to accommodate the nascent chain (Zhang et al., 2013). RpL4, rpL17 and rpL39 form the interior lining of the exit tunnel and can transduce signalling from interactions with the nascent peptide. Protein rpL39 is located at the external orifice of the tunnel and interacts with signal sequences of nascent peptides during co-translational insertion into the endoplasmic reticulum (Zhang et al., 2013). Co-translational nascent chain force measurements, inter-subunit fluorescence resonance energy transfer studies on single translating ribosomes, molecular dynamics simulation and cryo-electron microscopy studies have yielded data supporting the involvement of the 80S ribosome in protein folding (Nilsson et al., 2015). The sarcin-ricin loop on 60S subunit rRNA contains the GTPase - activating centre which binds incoming elongation factors (Shi et al., 2012). The (aminoacyl) A site is the binding location for charged tRNA, and can accommodate these tRNAs at a rate of ~10 s^{-1} (Bieling et al., 2006; Beringer and Rodnina, 2007). The (peptidyl) P site is the second binding site for tRNA and holds tRNAs which are linked to the growing polypeptide chain (Beringer and Rodnina, 2007). The 23S rRNA in the PTC catalyses the aminolysis of an ester bond between the alpha amino group of the A site amino acid and the carbonyl carbon of the P-site peptide amino acid (Simonović and Steitz, 2009). The (exit) E site contains deacylated tRNA, held here until dissociation from the ribosome. The mRNA binding groove forms upon binding of the 43S pre-initation complex to the large subunit (Ben-Shem et al., 2010). The groove wraps ~30 nucleotides of the mRNA at a time, and the channel is only wide enough to allow passage of single-stranded RNA (Yusupova et al., 2001). Additionally, the eukaryotic ribosome has been shown to be highly efficient at melting RNA secondary structures, and the putative ribosome-associated mRNA unwinding activity is thought to be (partially, at least) mediated by rpS30e, rpS3, rpS2 and rpS9 (Takyar et al., 2005; Namy et al., 2006; Wilson and Cate, 2012).

Besides the ribosome, other key components of the translational apparatus include transfer RNAs, which constitute 4-15% of all cellular RNA (Kirchner and Ignatova, 2014). tRNAs act as adaptor molecules that transfer the information encoded in mRNAs in a template-directed manner into a corresponding polypeptide chain (Kirchner and Ignatova, 2014). About 20% of all cellular transcription is devoted to producing tRNA and approximately three million new tRNAs are made during each cell division cycle (Grewal, 2014). tRNAs are exclusively transcribed by RNA polymerase III (Pol III), and several oncogenes are known to stimulate tRNA synthesis through regulation of this key enzyme. tRNA production involves the transcription of the initial RNA, followed by processing to remove the 5′ leader, trimming of the 3′ end, addition of a CCA, splicing of introns, modification of multiple nucleoside residues and export into the cytoplasm (Phizicky and Hopper, 2010). Aminoacyl tRNA synthetases

mediate aminoacylation, joining amino acids to specific acceptor tRNAs. Changes in tRNA levels can result in down-regulation or up-regulation of translation of specific mRNAs, such as GCN4, whose translation increases during amino acid starvation (Hinnebusch, 2005).

The ribosome interacts with a plethora of initiation, elongation and termination factors which regulate and facilitate the successive stages of translation (described in following sections). Other effectors of translation regulation include poly(A) binding proteins (PABPs), which are RNA binding proteins that can interact with 3′ UTRs and facilitate initiation as well as efficient translation of mRNAs (Goss and Kleiman, 2013). *cis*-acting sequences such as upstream ORFs or poly-A tails or secondary structures in a transcript can influence ribosome activity on a given transcript (Gebauer et al., 2012; Ma et al., 2012; Wethmar, 2014; Theil, 2015).

1.1.2 Initiation

The eukaryotic translation cycle (Figure 1.2) can be divided into four inter-connected phases: initiation, elongation, termination and ribosome recycling (Kapp and Lorsch, 2004). It is thought primarily to be regulated at the first of these stages (Jackson et al., 2010). Several independent methodologies have been used to estimate the average rate of translation initiation. Schwanhäusser et al. (2011) used pulse-chase labelling to calculate the rate of initiation as ~0.1-0.5 per second for the murine β-actin mRNA. Recent fluorescent nascent chain tracking (NCT) data showed initiation occurs every ~30 seconds on average & polysomes contain ~ 1 ribosome every 200-900 nucleotides (Morisaki et al., 2016). Wu et al. (2016) used single-molecule fluorescence recovery after photobleaching and single-molecule imaging of nascent peptides (SINAPS) in U2OS cells to calculate the average initiation rate as ~1.3-2 per minute. Initiation requires assembly of a pre-initiation complex (PIC) that proceeds to scan along a 5′ UTR until it encounters an appropriate start codon and can begin peptide bond catalysis. Eukaryotic initiation factor 1 (eIF1), eIF1A, eIF3 and eIF5 promote binding of Met-tRNA$_i$ with eIF2-GTP and the 40S subunit to form the 43S pre-initiation complex (PIC) (Hinnebusch, 2014). eIF1A cooperates with eIF1 to promote ribosome scanning and initiation codon selection. The 46 kDA eIF4A, 25 kDa eIF4E and 185 kDA eIF4G form the eIF4F complex which associates with the 5′ cap. eIF4A possesses DEAD-box ATPase and RNA helicase domains that melt secondary structures located in the 5′ UTR. The 69 kD eIF4B RNA binding protein enhances the helicase activity of eIF4A and the eIF4E polypeptide helps mediate the mRNA-ribosome binding step. eIF4G functions as a scaffolding polypeptide that binds eIF4E, eIF4A, PABP, SLIP1 and the

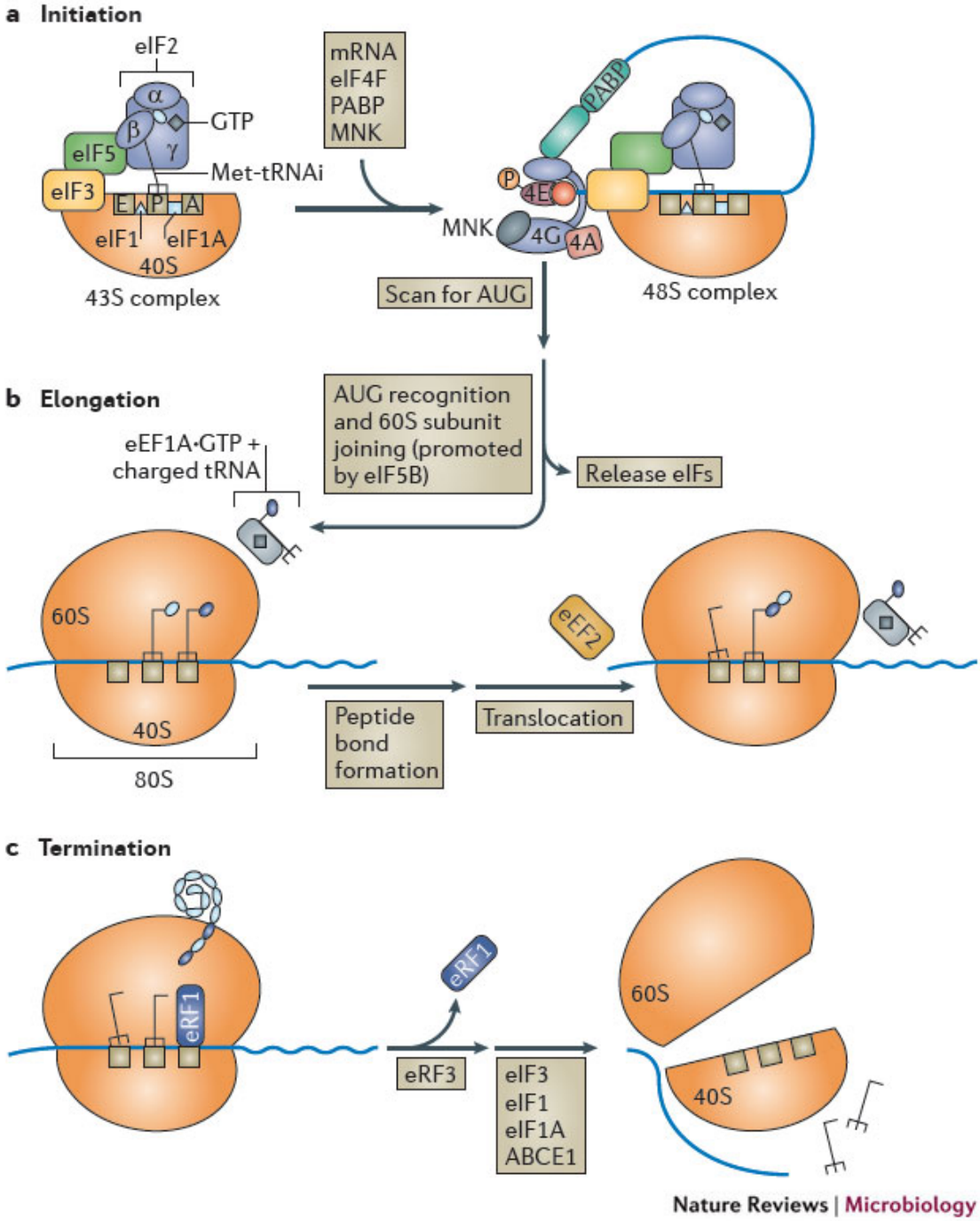


Figure 1.2 Overview of the eukaryotic translation cycle. During the initiation phase, the small subunit of the 80S ribosome loads onto the 5′ leader of a transcript along with several initiation factors and proceeds in the 3’ direction until an appropriate context start codon pairs with the initiation tRNA and the large subunit associates with the complex. Elongation factors facilitate the binding of charged tRNAs and GTP-GDP exchange. After the ribosome reaches an appropriate stop codon, the peptide chain is released from the exit tunnel and various factors mediate large and small subunit dissociation. Figure from Walsh and Mohr, (2011).

mRNA, as well as enhancing the helicase activity of eIF4A. Additionally, interactions between the eIF4G component and 3′ UTR associated PABP are thought to mediate circularization of transcripts to facilitate efficient translation (Wells et al., 2016). The association of the cap-binding eIF4E protein is thought to be the rate limiting step of (cap-dependent) initiation (Mochizuki et al., 2005). One of the many functions of eIF3 is to prevent the 60S subunit from associating with the PIC until a proper start codon has been matched.

The 43S pre-initiation complex (PIC) scans in a 5′-3′ direction along the transcript, and is thought to protect 40-70 nt of the mRNA (Lazarowitz and Robertson, 1977; Pisarev et al., 2008; Dunn et al., 2013). After an appropriate start codon base pairs with the anticodon loop of the P-site Met-tRNA^{Met}, the 48S complex forms. eIF5/eIF5B induce the hydrolysis of eIF2-bound GTP, displacement of the various initiation factors and association of the 60S subunit. The likelihood of a full ribosome assembling on a given start codon depends greatly on the local sequence context, with the presence of a Kozak consensus sequence giving optimal initiation - GCC(A/G)CCAUGG (start codon underlined), with a purine at the –3 and a G at the +4 positions (relative to the A of the AUG codon, which is designated +1) (Kozak, 1991; Jackson et al., 2010). eIF1 mediates high-fidelity recognition of an appropriate initiation codon (Sonenberg and Hinnebusch, 2009). The importance of the Kozak context is exemplified by the phenomenon of leaky scanning, where a pre-initiation complex may continue moving along a transcript past a potential (AUG) start codon due to the poor initiation context surrounding that codon (Andrews and Rothnagel, 2014). Reinitiation after uORF-translation is not fully understood, though current models posit it as an inefficient mechanism (Barbosa et al., 2013). Several initiation factors need to remain associated with the ribosome during translation and after termination takes place so that reinitiation can occur (Pöyry et al., 2004). Efficient reinitiation has been shown to occur only if an eIF4F complex was involved in mediating uORF initiation (as opposed to when an IRES is involved in facilitating initiation). These results suggest that resumption of scanning may depend on the interaction between eIF4F (or the eIF4G central domain) and the ribosome being maintained while the ribosome translates a short uORF.

1.1.3 Elongation

After initiation, a fully assembled 80S ribosome is present on the transcript with the anticodon of Met-RNA$_i$ in the P site base-paired with the start codon (Dever and Green, 2012). The second codon of the ORF is situated in the A site. The GTPase eEF1A binds an amino-acyl-tRNA, forming an eEF1A-GTP-aminoacyl-tRNA ternary complex that delivers the

adaptor to the A site (Sasikumar et al., 2012). Codon:anticodon pairing induces GTP hydrolysis by eEF1A and results in eEF1A-GDP dissociating and tRNA accommodation. The guanine nucleotide exchange factor eEF1B catalyses guanine nucleotide exchange on eEF1A to recycle the elongation factor. Once the charged tRNA is accommodated in the A site, the PTC catalyses peptide bond formation with the peptidyl tRNA. The translocation step involves a ratcheting motion between the ribosome subunits that produces a transient hybrid state and binding of the GTP/eEF2. Hydrolysis of GTP is thought to allow rearrangements that complete the movement of tRNA and mRNA, followed by locking of the subunits and release of inorganic phosphate and eEF2 (Dever and Green, 2012). At this point, the E site contains a deacylated tRNA and the A site is empty, and the ribosome can repeat the process to continue synthesizing the peptide. Translation elongation is thought to be a high-fidelity process, with the 80S ribosome only making $1/10^4$ mis-incorporations (Wohlgemuth et al., 2011). The high level of accuracy of ribosome decoding reactions, is partially due to preferential rejection of incorrect aa-tRNAs in a step known as 'proofreading' that occurs after GTP-hydrolysis (Wohlgemuth et al., 2011).

Various methods have been employed to calculate the average rate of eukaryotic elongation. Von der Haar (2008) constructed a computational model of translation kinetics based on several curated transcript abundance, protein abundance and protein half-life measurements to estimate the elongation rate as 32.6 codons/second. Ribosome profiling in murine embryonic stem cells yielded an estimate of 5.6 codons/sec (Ingolia et al., 2011) while ribosome profiling of murine 17 Cl-1 cells gave 4.6 codons per second (Irigoyen et al., 2016). Recent advances in single-molecule translation imaging techniques have yielded estimates that corroborate the ribosome profiling estimates. Several groups independently developed an *in vivo* imaging technique where cells are transfected with reporter transcripts encoding SunTag peptides followed by a gene of interest along with a second construct that expressed a GFP-tagged antibody against the SunTag epitopes (Wang et al., 2016; Wu et al., 2016; Yan et al., 2016). This approach was used to visualize translation of single mRNAs *in vivo* for over an hour and to estimate the rate of ribosome translocation (~3.5 ± 1.1 codons/second), though the authors noted considerable variability in elongation rates even amongst different copies of the same transcript isoforms. One group developed an analogous system that used FLAG tags and obtained similar values for the elongation rate (Morisaki et al., 2016).

1.1.4 Termination

When a UAA, UGA or UAG stop codon enters the A site, a ternary complex consisting of eukaryotic release factor 1 (eRF1), eRF3 and GTP binds to the A site in a pre-accommodated state (Jackson et al., 2012). After GTP hydrolysis occurs and eRF3 is released, the ATPase ABCE1 binds and induces an active configuration where eRF1 facilitates peptidyl-tRNA hydrolysis and release of the polypeptide. The small and large subunits of the ribosome can then dissociate (Dever and Green, 2012). After termination, ribosomes are recycled for following rounds of translation, though the mechanism behind this phenomenon is not fully understood (Dever and Green, 2012). While bacterial ribosomes utilize a tRNA-like molecule called the ribosome recycling factor (RRF) to split 70S ribosomes and release mRNA templates after termination (in conjuction with EF-G), no similar protein has been identified in eukaryotes or archaea (Janosi et al., 1994; Nürenberg and Tampé, 2013; Iwakura et al., 2017). Circularization of transcripts, mediated through RNA binding proteins such as PABP, is thought to promote efficient recycling of ribosomes for repeated translation of the same message (Jackson et al., 2010). In some scenarios, such as during depletion of eRF1, translational readthrough occurs and ribosomes continue elongation past an in-frame stop codon resulting in a C-terminal extension protein (Dinman and Berry, 2007). The UGA stop codon in some transcripts that contain specific RNA structures can pair with the seryl-tRNA$^{(Sec)}$, resulting in incorporation of a selenocysteine residue (Schmidt and Simonović, 2012). The 40S subunit may occasionally remain in contact with a transcript after dissociation of the 60S subunit and resume scanning until it encounters another start codon – this is particularly relevant to cases of short ORFs in 5′ UTRs (Andrews and Rothnagel, 2014). In these scenarios, the small subunit is initially unable to resume translation due a lack of a eIF2 ternary complex but is still scanning capable (Jackson et al., 2010).

1.1.5 Comparison of various ribosomes and translation systems

Though the discussion thus far has centred on cytoplasmic, mammalian or yeast ribosomes, comparisons will be made with other types of ribosomes and protein synthesis mechanisms to highlight key similarities and differences. Prokaryotic ribosomes are 50% smaller and only about one-third of the 80 to 90 eukaryotic ribosomal proteins have bacterial counterparts (Ramakrishnan, 2011). The 70S prokaryotic ribosome consists of 50S and 30S subunits, and lack 5.8S rRNA. Many core functionalities such as peptidyl transfer and translocation are conserved between bacteria and eukaryotes. However, eukaryotic translation is more intricate and highly regulated (Sonenberg and Hinnebusch, 2009), especially during

initiation, which requires about a dozen initiation factors rather than three in prokaryotes (Berg et al., 2012). Eukaryotes have various translation surveillance and quality control mechanisms such as Nonsense-mediated mRNA decay (NMD) that have not been identified in prokaryotes (Ramakrishnan, 2011). In prokaryotes, the initiating amino acid is a formylated methione (Gualerzi and Pon, 2015). Prokaryotic elongation is in many regards similar to eukaryotic elongation - the elongation factors EF1α and EF1βγ are the counterparts of prokaryotic EF-Tu and EF-Ts (Rodnina and Wintermeyer, 2009). Eukaryotes generally use one termination factor that recognizes all three stop codons, while prokaryotes have one factor that recognizes UAA/UAG and one factor that recognizes UAA/UGA stop codons (Cridge et al., 2006). As mentioned above, bacterial ribosomes utilize the RRF to facilitate recycling post-termination, while a similar factor has not been identified for eukaryotes (Nürenberg and Tampé, 2013).

Several features of mitochondrial translation in eukaryotes have closer resemblance to prokaryotic translation rather than to that of eukaryotic cytoplasmic translation. Mitochondrial ribosomes, which translate the protein coding transcripts produced by 13 of the 37 mitochondria genes, are structurally more similar to bacterial than eukaryotic ribosomes (Sharma et al., 2003). The vertebrate mitochondrial genetic code has some variations on the cytosolic one, including the use of UGA to encode tryptophan instead of stop, the presence of only 22 distinct tRNAs, the use of both AUG and AUA as initiation codons, and the lack of amino acid assignment for the AGG and AGA arginine codons (Smits et al., 2010). Mitochondrial transcripts also lack UTRs and 5′ cap structures.

The chloroplast ribosome in plants is also structurally more related to prokaryotic rather than 80S ribosomes, though it also utilizes a number of unique proteins, some of which are thought to mediate light-sensitive translation of certain transcripts (Manuell et al., 2007). Unspliced chloroplast transcripts are located in the same compartment as chloroplast ribosomes (unlike pre-mRNA encoded by genes in the nucleus). Chloroplast ribosomes have been observed translating on unspliced transcripts and terminating within introns; many chloroplast transcripts are also polycistronic (Zoschke et al., 2013). Translation of some chloroplast ORFs requires the involvement of specific *trans*-acting factors (Zoschke et al., 2013; Chotewutmontri and Barkan, 2016).

Bacterial transcripts have been known to be poly-cistronic for quite some time, though eukaryotic (animal) mRNAs were thought to be exclusively mono-cistronic until recently (Mouilleron et al., 2016). The *GNAS* (G-protein α-subunit) transcript contains two reading frames and has been shown to produce two structurally unrelated proteins, XLαs and ALEX (Xu et al., 2010). Raj et al. (2016) mined ribosome profiling and mass spectrometry datasets to

infer several hundred example of mammalian poly-cistronic RNAs, noting that 40% of highly supported candidates showed negative correlations in the translation levels of their two coding sequences. Many newly identified secondary ORFs are short (<200 aa), which has hindered their detection in previous *in silico* annotations of eukaryotic transcripts (Dinger et al., 2008).

1.1.6 Noncanonical translation mechanisms

While the previously described mechanism of translation represents a model that describes translation of the majority of eukaryotic transcripts under 'typical' conditions of cellular homeostasis, many exceptions to this model have been documented, and it continues evolve as new methodologies uncover previously undiscovered mechanisms of translation control (

Figure 1.3). Many examples of non-canonical translation were first characterized in viruses.

Viruses are, by definition, obligate intracellular parasites. The production of new viral proteins is universally dependent on the cellular translation apparatus (Firth and Brierley, 2012). Many general features about translation were discovered in the context of virus infection through understanding of how these pathogens cause or use deviations from "standard" protein synthesis. For example, investigations of vesicular stomatitis virus (VSV) infected cells revealed that while all VSV transcripts required functional rpL40 on the surface of 80S ribosomes to produce their proteins, only 7% of cellular transcripts (mostly stress response genes) required this component for translation, suggesting that certain ribosome components act as regulators for select subsets of mRNAs (Lee et al., 2013). These infectious agents have to overcome various challenges imposed by the physical constraints of capsid diameter on their genome sizes and structural idiosyncrasies in various viral transcripts. Many viruses have evolved atypical translation mechanisms in order to overcome these obstacles.

1.1.7 Alternative initiation

While cellular initiation usually begins with eIF4F recognition of the 5′ 7-methylguanylate cap, several viruses have been shown to forego this mechanism. The use of internal ribosome entry sites (IRES) to enable ribosome loading in the absence of a 5′ cap and/or to induce initiation at a secondary AUG was first described in picornaviruses (Jang et al., 1988; Pelletier and Sonenberg, 1988; Firth and Brierley, 2012), and has been found in hepatitis C

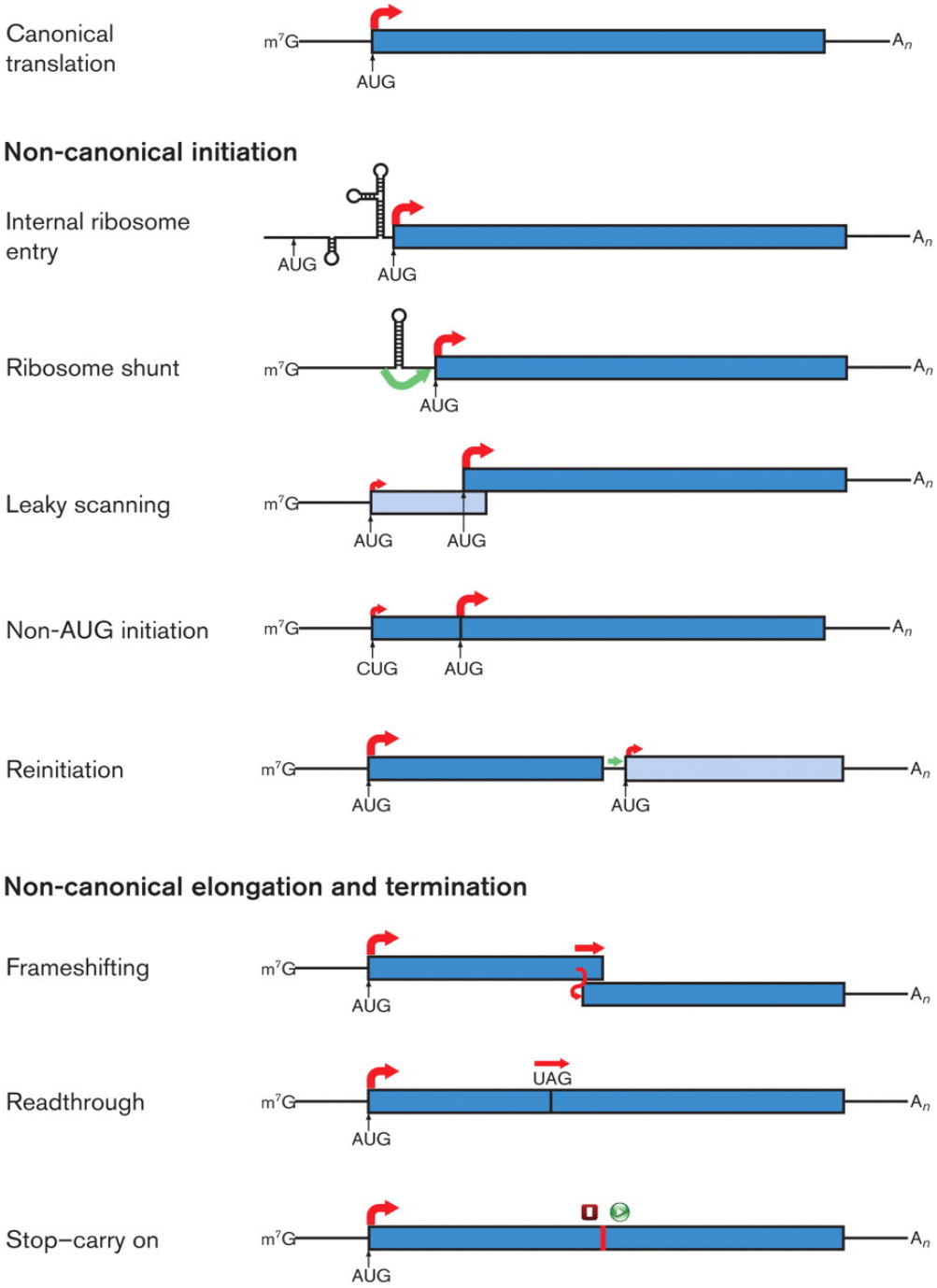


Figure 1.3 Non-canonical initiation, elongation and termination mechanisms. Various viral, eukaryotic and prokaryotic mechanisms allow for ribosomes to initiate at alternative sites that produce N-terminally extended or truncated isoforms, or proteins from completely different ORFs. Re-initiation in the context of transcripts with uORFs functions as a stochastic regulatory mechanism, such as in the translation of the yeast GCN4 protein. Frameshifting and readthrough result in protein variants with different C-termini, while Stop-carry on results in the production of two separate proteins from one ORF when the ribosome prematurely releases the first peptide and then synthesizes a second peptide. Figure from Firth and Brierley, (2012). Bypassing and StopGo translation are not shown in figure.

hepacivirus, Theiler's murine encephalomyelitis virus, foot-and-mouth disease virus, dicistroviruses and other viruses. Additionally, some viruses are able to aid translation of their transcripts by interfering with various components of cap dependent translation: polioviruses cleave eIF4G and influenza virus decreases eIF4E phosphorylation to reduce ribosome recruitment to cellular mRNAs (Walsh et al., 2013). Leaky scanning has been found to occur in the context of several RNA virus infections. In this process, a significant proportion of ribosomes fail to initiate at the first AUG and continue scanning until reaching a secondary AUG to produce a variety of proteins with different N-terminal ends. Ernst and Shatkin (1985) identified one of the first examples of leaky scanning in segment S1 of the mammalian orthoreovirus. Fuller et al. (1983) found that the NSs protein of orthobunyaviruses is translated from an ORF that overlaps the 5′ portion of the viral nucleocapsid encoding ORF via leaky scanning. While translation by 80S ribosomes is thought to primarily occur starting at AUG Curran and Kolakofsky (1988) described one of the first cases of non-AUG initiation in Sendai respirovirus, where an ACG codon is used to commence synthesis of a functional, N-terminally extended variant of the viral C protein. While pre-initiation ribosomes generally process in a step-wise, 5′-3′ direction along most transcripts, in some cases they can be shunted to a downstream segment in a scanning-independent mechanism (described in Futterer et al., (1993) and Ryabova et al., (2006)).

1.1.8 Alternative elongation

While translation is primarily regulated at the level of initiation, there are several mechanisms through which specific transcripts can be translated in alternative ways to yield different proteins. Several of these processes were originally discovered in RNA viruses, which have unique needs for regulating gene expression and spatial constraints on their genomes that necessitate maximizing the coding capacity of their nucleic acids.

The various categories of recoding mechanisms include ribosomal frameshifting, bypassing, stop-codon reassignment, stop codon readthrough and stop-go (Atkins et al., 2010; Firth and Brierley, 2012). Jacks and Varmus (1985) identified the first example of programmed frameshifting in Rous sarcoma virus during expression of its Gag-Pol polyprotein. During ribosomal frameshifting a proportion of elongating ribosomes are directed into a different reading frame by shifting forwards or backwards 1 or 2 nt in response to specific signals in the mRNA, namely a heptanucleotide slippery sequence and a stimulatory RNA secondary structure (Atkins et al., 2010). Frameshifting has since been found to occur in a variety of other

viruses including clinically relevant pathogens (such as HIV). The first cellular frameshifting sample was found in the *E. coli dnaX* gene (Flower and McHenry, 1990; Tsuchihashi and Kornberg, 1990; Yan et al., 2015), while the first mammalian example was identified in the mammalian genes for ornithine decarboxylase antizymes (OAZ1, OAZ2, OAZ3), which regulate polyamine levels (Matsufuji et al., 1995; Atkins et al., 2010). Frameshifting occurs in the mammalian genes for embryonic development protein PEG10 (Manktelow et al., 2005), the tumour-development linked Ma3 gene (Wills et al., 2006) and has been suggested to occur in the chemokine receptor CCR5 that is used by HIV for cell entry (Belew et al., 2014),. Napthine et al. (2016) documented the first instance of *trans*-activation of frameshifting (in the absence of a stimulatory RNA structure) in porcine reproductive and respiratory syndrome virus (PRSSV). During PRSSV infection, a complex of viral nsp1β and host poly(C) binding protein can *trans*-stimulate frameshifting to produce a transframe fusion protein (nsp2TF).

1.1.9 Modified termination

Though stop-codon redefinition to encode amino acids such as selenocysteine has been documented in a limited number of viruses (such as the Molluscum contagiosum poxvirus glutathione peroxidase (Shisler et al., 1998)), there are numerous instances of viral translational readthrough. The first example of stop codon readthrough was identified in the RNA bacteriophage Qβ, resulting in the production of a C-terminally extended coat protein which is essential for progeny virion formation (Hirsh, 1971; Beier and Grimm, 2001). Csibra et al. (2014) showed that decreasing readthrough efficiency even modestly in murine leukemia virus (MuLV) substantially abrogated viral replication. The human *OPRL1*, *PRK1*, *MAPK10* and *AQP4* genes were found to contain functional readthrough signals as observed using reporter constructs and western blotting (Loughran et al., 2014). Programmed readthrough has recently been found to occur in several hundred metazoan ORFs and in many cases, results in C-terminal extensions that contain localization signals (Dunn et al., 2013).

In bypassing, P-site tRNA anticodon dissociation, mRNA slippage and repairing at a non-overlapping codon result in the synthesis of a single protein from two separated ORFs (Atkins et al., 2010). The first example of programmed bypassing was identified between the two ORFs of bacteriophage T4 gene *60*, which are separated by 50 nt (Huang et al., 1988; Weiss et al., 1990).

StopGo translation involves a mechanism that directs co-translational separation of the peptide chain by preventing peptide-bond formation at a specific site (Brown and Ryan, 2010;

Sharma et al., 2012). A nascent peptide fragment interacts with part of the ribosome exit tunnel, inducing ribosome pausing, stop-codon independent protein release and continued translation to yield a second protein. Various positive-stranded mammalian RNA viruses and double-stranded insect RNA viruses have been shown to utilise StopGO translation (Donnelly et al., 2001; De Felipe et al., 2003).

1.2 **Ribosome Profiling**

The advent of high throughput next generation sequencing has enabled the assembly of vast numbers of genomic sequences and identification of population and disease-related DNA sequence variation (Buermans and den Dunnen, 2014; Goodwin et al., 2016). The use of reverse-transcriptase and nucleic acid linkers has extended the reach of this technique to the transcriptome by enabling the sequencing of the mRNA content of cells via RNA-Seq (Mortazavi et al., 2008). Simultaneous advances in mass spectrometry instrumentation and informatics have allowed a deeper understanding of various proteomes (Altelaar et al., 2013; Larance and Lamond, 2015). However, the aforementioned methodologies have a teleological gap in that they do not provide information about the temporal-spatial kinetics of new protein synthesis (merely its accumulated abundance) and use of various mechanisms to expand the protein-encoding capacity of transcripts. The interface between the domains of transcriptomics and proteomics has been dubbed ‘translatomics’ (Kuersten et al., 2013; Neuhaus et al., 2016; Yang et al., 2016). Polysome profiling was the first technique to allow high-throughput investigation of the translatome, though it has low quantitative or spatial resolution for specific mRNA species (Ingolia, 2016). Steitz (1969) showed that translation competent ribosomes protect a fairly uniform ~30 nt portion of the attached transcript.

Ingolia’s seminal contribution was the coupling of protected fragments with next-generation sequencing technology to produce a snapshot of global translation *in vivo* with high quantitative and spatial resolution. This methodology – dubbed ‘ribosome profiling’ - has helped bridge the gap between transcriptomics and proteomics, and enabled a large number of new analyses of ribosome activity.

1.2.1 Experimental workflow of ribosome profiling

In one version of the protocol, sells are lysed in cryogenic conditions, often in the presence of a translation inhibitor such as cycloheximide or harringtonine and treated with a ribonuclease to degrade portions of the mRNA transcript that are not protected by translating ribosomes (Figure 1.4). Ribosomes and their associated footprints are pelleted through a sucrose cushion, and the resulting ribosome protected footprints (RPFs) are precipitated by hot phenol/chloroform extraction. These fragments are gel purified, 3′ dephosphorylated and 5′ phosphorylated. Illumina-compatible 5′ and 3′ adaptors are ligated, after which the reads are amplified by RT-PCR. ~80% of cellular RNA in growing mammalian cells is rRNA (Lodish et al., 2000), and rRNA contamination decreases the amount of informative sequence data obtained in a sequencing experiment (Ingolia et al., 2012). Prior to the amplification step, rRNA contamination can be reduced through hybridization to biotinylated antisense-strand oligonucleotides or through use of a duplex specific nuclease (Chung et al., 2015).

An RNA-Seq library is typically prepared in parallel to allow for transcript normalization and aid in isoform detection. The resulting RiboSeq and RNASeq libraries are sequenced on an Illumina sequence-by-synthesis platform. A typical RiboSeq library sequenced on a NextSeq platform yields several gigabytes of data, representing between 10^6 and 10^7 reads. In a commonly used protocol for ribosome profiling (Ingolia et al., 2012), the raw sequence data are then computationally clipped and trimmed to remove adaptor and low quality sequences. The trimmed data are aligned to an rRNA index using a short-read aligner such as bowtie (Langmead, 2010). The remaining reads are mapped to a given genome with a corresponding transcriptome annotation using a splice-aware short read aligner such as TopHat (Trapnell et al., 2009). Once a binary alignment file has been produced (Li et al., 2009), various downstream analyses including differential expression, read distribution visualization and sequence enrichment can be performed.

1.2.2 Applications of ribosome profiling

Since its introduction in 2009, ribosome profiling has been used in scores of investigations to investigate various translation phenomenon in a myriad of organisms (Figure 1.5), including sub-cellular organelles such as the ER, mitochondria and chloroplasts. The technique has also been applied to several pathogens, including the DNA virus bacteriophage lambda (Liu et al., 2013b), the vaccinia poxvirus (Yang et al., 2015), human cytomegalovirus (HCMV) (Stern-Ginossar et al., 2012) and Kaposi's sarcoma-associated herpesvirus (Arias et al., 2014), as well as the RNA virus, mouse hepatitis virus (MHV) (Irigoyen et al., 2016) and

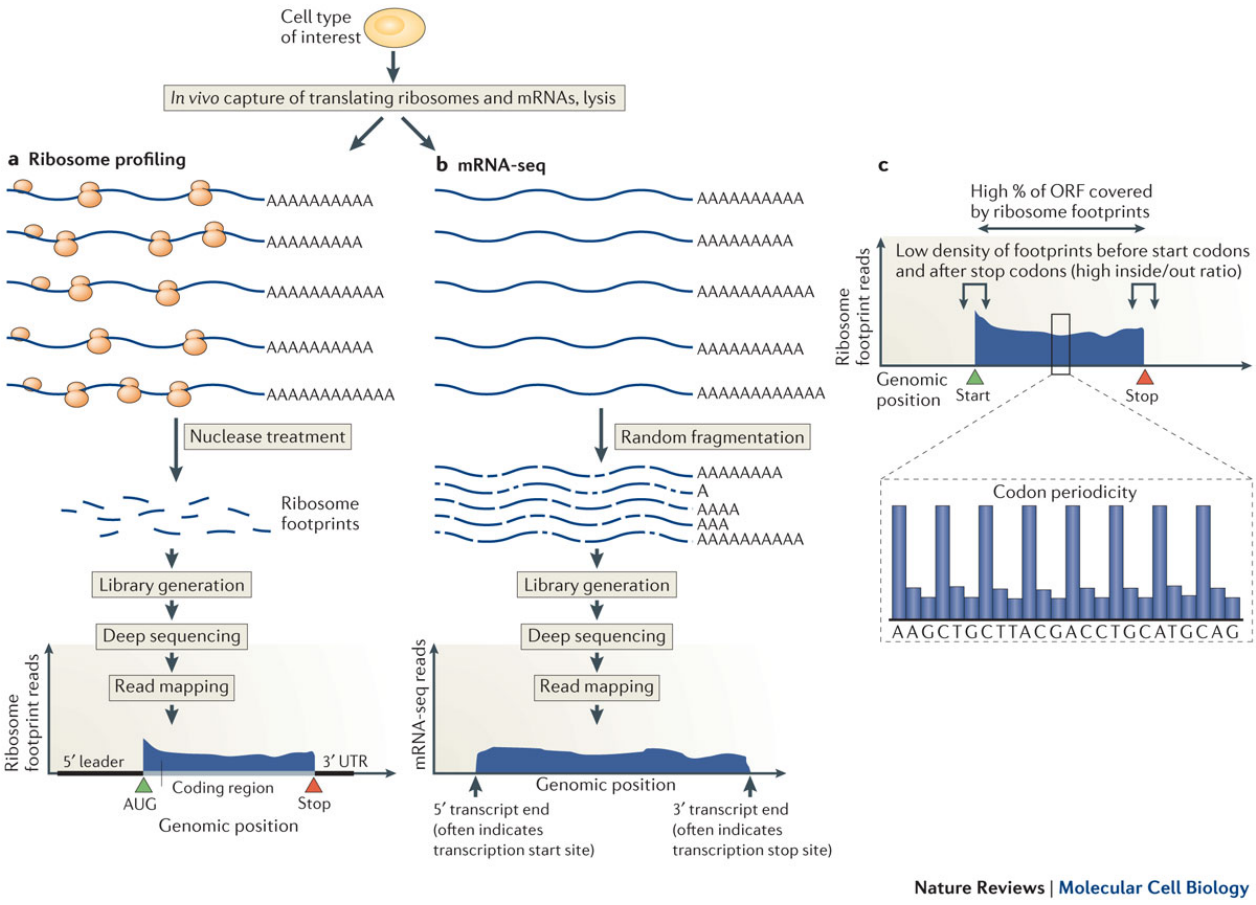


Figure 1.4 Overview of ribosome profiling. During RiboSeq library preparation, cells are cryogenically treated and lysed; some studies also add chemical inhibitors of translation. Unprotected RNA is digested by nuclease treatment and RPFs are isolated, size selected and converted into an adaptor-linked cDNA library which is then sequenced. RNASeq libraries are prepared in a similar fashion, except no monosome isolation is performed and transcripts are fragmented by alkaline hydrolysis or other means. RiboSeq libraries typically display a high ratio of reads aligning to coding versus non-coding segments. Figure adapted from Brar and Weissman, (2015).

the parasitic protozoan *Trypanosoma brucei* (Jensen et al., 2014; Vasquez et al., 2014). Profiling has been used to investigate many experimental objectives, which can be broadly grouped into the annotation of translated sequences, the characterisation of mechanisms of protein synthesis, and the measurement of translational control of gene expression (Ingolia, 2016). A distinctive feature of profiling datasets is the 'phasing' of RPFs – the tendency for read alignments to have a three nucleotide periodicity due to the triplet nature of genetic decoding and the uniform trimming produced by (certain) nucleases. Thus profiling can also be informative about the utilisation of various ORFs, including short or overlapping ORFs.

1.2.3 Annotation of ORFs

Purely-sequence dependent methods of predicting coding sequences have struggled to reliably annotate short ORFs (Dinger et al., 2008). Many of these approaches rely on evidence of evolutionary conservation and are thus unsuited to recently evolved or species-specific genetic elements, which may be over-represented in small ORFs (Carvunis, 2012; Reinhardt and Jones, 2013). Ribosome profiling has been used to identify novel translated regions and several algorithms have been developed to mine RiboSeq data for evidence of translational activity (Table 1.1, Table 1.2). Several studies have identified highly-periodic, AUG-initiated, RPF distributions on transcripts that were previously labelled as long-non-coding RNAs due to a lack of conserved protein-coding potential (Ingolia et al., 2011; Bazzini et al., 2012; Brar et al., 2012; Stern-Ginossar et al., 2012). RPFs originating from some of these 'non-coding RNAs' have been corroborated by detection of the predicted peptides via mass spectrometry (Stern-Ginossar et al., 2012; Slavoff, 2013; Fields et al., 2015). Multiple profiling studies have also found evidence of translation of the 5′ leaders of annotated coding transcripts (Ingolia et al., 2009; Lee et al., 2012; Andreev et al., 2015a; Calviello et al., 2015; Tzani et al., 2016). Some of these upstream ORFs (uORFs) may have regulatory functions, such as to occupy scanning ribosomes and to decrease the translation of downstream ORFs. Near-cognate initiation codons (CUG, GUG, UUG, ACG) were found to be ~four times more common than AUGs in newly detected uORFs in mouse embryonic stem cells (Jackson and Standart, 2015). Raj et al. (2016) developed the riboHMM algorithm to mine ribosome footprint data to infer translated sequences and applied it to human lymphoblastoid cell lines; they identified 7273 novel coding sequences and 2442 translated uORFs. 40% of bicistronic transcripts showed a negative correlation in the translation levels of two coding sequences, lending additional support to the hypothesized regulatory function of these elements. Additionally, 14% of the predicted uORFs

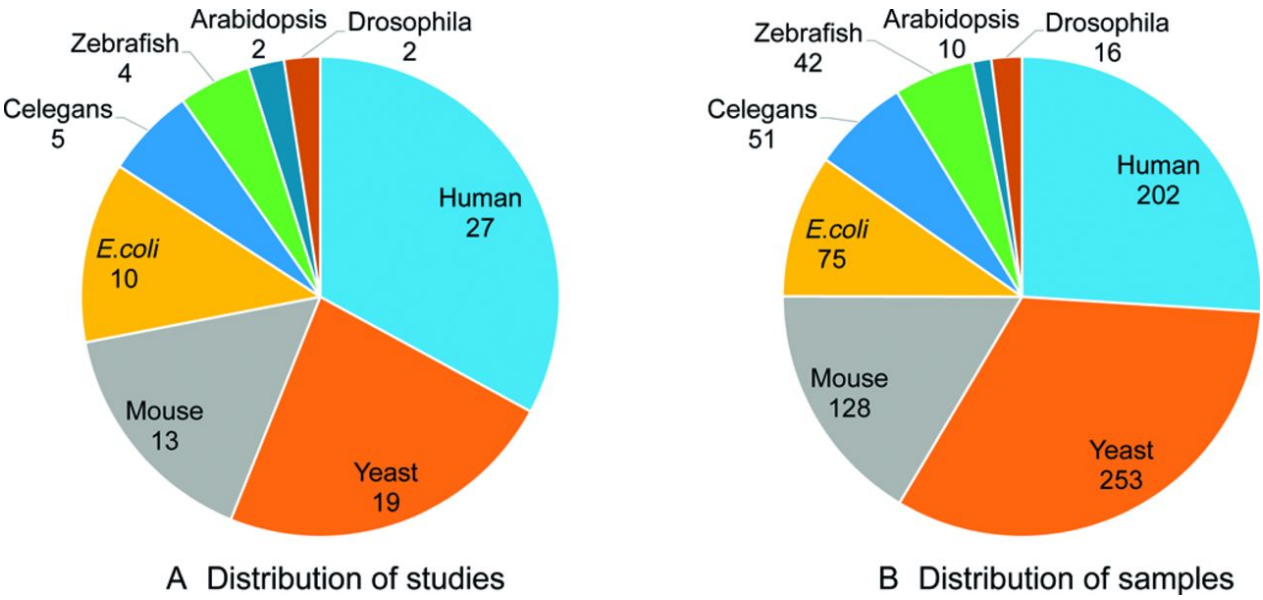


Figure 1.5 Published ribosome profiling studies by organism. (A) Number of RiboSeq studies published on cells from each species. (B) Number of individual RiboSeq replicates published per species, as of January, 2016. In addition to cellular studies, there have been profiling reports on HCMV, bacteriophage lambda, the vaccinia poxvirus, Kaposi's sarcoma-associated herpesvirus, murine hepatitis virus and *Trypanosoma brucei*. Figure from Xie et al., (2016).

Database	Data Collected	Reference
TISdb	Translation initiation sites	Wan and Qian, (2014)
GWIPS-viz	RPF & RNASeq genome browser	Michel et al., (2014)
sORFs.org	Short ORF annotations	Olexiouk et al., (2015)
RPFdb	RPF genome browser; expression measurements	Xie et al., (2016)

Table 1.1 List of ribosome profiling databases. Table adapted from Ingolia, (2016)

Algorithm or Metric	Input Data	Output Classification	Reference
Periodicity transition score	elongating ribosome frame	Dual-coding regions	Michel et al. (2012)
Translated ORF classifier	elongating ribosome density	CDS ORF, 5′ UTR ORF, 3′ UTR ORF	Chew et al. (2013)
Ribosome release score	elongating ribosome density	CDS-like	Guttman et al. (2013)
N/A	elongating ribosome density	stop codon read-through	Dunn et al. (2013)
Change point analysis	elongating ribosome occupancy	novel isoforms; alternate frames; drop-off	Zupanic et al. (2014)
FLOSS	footprint length	true ribosome occupancy	Ingolia et al. (2014)
ORF score	elongating ribosome frame	short ORFs	Bazzini et al. (2014)
RiboSeqR	elongating ribosome frame & density	short ORFs	Hardcastle (2014)
PROTEOFORMER	elongating ribosome density; mass spectrometry	short ORFs; novel isoforms	Crappé et al. (2015)
ORF-RATER	elongating ribosome frame; footprint length; Harr/LTM initiation	short ORFs; novel isoforms	Fields et al. (2015)
RibORF	elongating ribosome frame; elongating ribosome evenness	short ORFs	Ji et al. (2015)
RiboHMM	abundance & periodicity of RPFs	uORFs, dual-coding, CDS in ncRNA	Raj et al. (2016)
RiboTaper	elongating ribosome frame	short ORFs; novel isoforms	Calviello et al. (2016)

Table 1.2 Algorithms for annotating translated sequences with Ribosome Profiling data. Numerous algorithms have been created for predicting short ORFs, frame-shift events, stop codon read through events, non-AUG initiation sites and short ORFs. Figure adapted from Ingolia, (2016).

in this study were validated by mass spectrometry. Other integrative analyses also detected high levels of novel translated sequences and partial confirmation by proteomics (Calviello et al., 2015; Fields et al., 2015).

There has been some debate regarding the role of putative small ORFs in 5′ leaders or previously unannotated transcripts (Guttman et al., 2013; Brar and Weissman, 2015; Ingolia, 2016). Proteomics has long served as an essential component of determining the protein-coding potential of genes, though it is known to be constrained by several experimental limitations. Mass spectrometry struggles to detect proteins expressed at low levels, and many of the newly detecting ORFs encode short peptides that are probably unstable in the cells, which may explain their low detection rate (Ma, 2014). Specific modifications in mass spectrometry protocols are being developed to allow for higher sensitivity towards small proteins (Slavoff et al., 2013; Dallas et al., 2015) and may in the future allow for more extensive validation of the various predicted short ORFs.

While it is likely that short ORFs have a heterogeneous mix of functions, there are numerous pieces of evidence to support the argument that at least some of these sequences have biologically functional, protein-encoding activity (Saghatelian and Couso, 2015). The *Drosophila tal/pri* gene encodes an 11 amino acid peptide whose mutation results in truncated limbs and the ~50 amino acid Zebrafish Toddler protein is involved in cell motility during embryogenesis (Kondo et al., 2007; Pauli et al., 2014). Profiling was used to detect translational activity on a short ORF on the β2.7 RNA of HCMV, which was previously thought to be non-coding (Reeves et al., 2007; Stern-Ginossar et al., 2012). Accumulation of the peptide was confirmed during HCMV infection, and was also detected in serum from HCMV-positive patients. It has been hypothesized that short ORFs may be displayed on cell surfaces via MHC I for immune surveillance of viral infection, cancer-associated mutations or autoimmune reactivity (Stern-Ginossar et al., 2012; Starck et al., 2016).

1.2.4 Detection of alternative initiation and termination sites

In addition to detecting completely novel translated sequences *in vivo*, ribosome profiling (particularly with initiation profiling using harringtonine or other analogous drugs) has also been utilized to detect novel isoforms of previously known proteins (Ingolia et al., 2011; Fritsch, 2012; Lee et al., 2012). Many of these novel proteins have been confirmed by proteomics (Fournier, 2012; Menschaert et al., 2013). These isoforms can be produced by use of an alternative upstream or downstream start codon, resulting in synthesis of an N-terminally

extended or truncated protein. These variations can affect the cellular localization of a synthesized protein (including its secretory capacity) or enable different and even opposing functions. For example, Brubaker et al. (2014) used profiling to show that a single transcript encodes a regulator of antiviral innate immunity, the RIG-I like receptor adaptor protein MAVS, which is produced in a full-length version that stimulates interferon signalling and a N-terminally truncated isoform that uses a downstream AUG to encode a variant that interacts with the full-length isoform but lacks key signalling domains and thus downregulates interferon production. Similarly, Williams and colleagues (2014) showed that fumarate reductase Osm1 also utilizes alternative initiation sites that encode distinct targeting signals, one of which directs the enzyme to the mitochondria and one which directs it to the ER.

Profiling has also been used to identify examples of programmed translational readthrough. Dunn et al. (2013) identified 350 examples of species-specific readthrough in *Drosophila*, many of which they were able to validate as encoding C-terminal targeting signals via reporter constructs and fluorescence microscopy. Schueren (2014) computationally mined profiling data and experimentally confirmed that >1.6% of mammalian lactate dehydrogenase B is synthesized via a readthrough event that produces a peroxisomal targeting signal that directs the enzyme to peroxisomes.

1.2.5 Study of translation mechanisms

Besides its applications in sequence annotation, ribosome profiling has been used to elucidate several details regarding the general mechanism of translation. As the ribosome ratchets along an mRNA during elongation, it assumes different conformations, one which was shown to protect ~28-30 nt while another elongation conformation protects 20-22 nt (Lareau et al., 2014).

Rooijers et al. (2013) showed that the size distribution of wild-type mitochondrial ribosome protected footprints (RPFs) follows a bimodal distribution peaking at 27 and 33 nt, unlike the cytoplasmic RPFs, which are typically 30 nt in length. This was similar to the RPF length distribution seen in bacteria (O'Connor et al., 2013). Guydosh and Green (2014) used ribosome profiling along with Dom34 depletion to demonstrate that the protein acts as a rescue factor which has an important role in freeing ribosomes from truncated transcripts and 3′ UTRs and that ribosomes are not always automatically released following stop codon recognition. Young et al. (2015) similarly used profiling to study the effects of ribosome recycling factor Rli1, observing that when the protein was depleted, 80S ribosomes accumulate at stop codons

and in the 3′ UTR. Some of these ribosomes reinitiated translation, a result that was corroborated by western blots and mass spectrometry analyses. Follow-up analyses indicated that simultaneous Dom34 and Rli depletion results in a dramatic increase in ribosomes in the 3′ UTR, suggesting that the function of Dom34 is to clear the majority of unrecycled ribosomes. While it has been previously stated that translation is primarily regulated at the level of initiation, several studies used RiboSeq to investigate the effects of various stressors on translation and found that the majority of changes were at the level of elongation. Gerashchenko et al (2012) used hydrogen peroxide to induce oxidative stress in yeast, and found that translational arrest occurred systematically ~50 codons downstream of transcript initiation sites. Shalgi et al. (2013) similarly found that heat shock induces translational arrest ~65 codons downstream of start sites in murine cells. Transcripts that were found to be especially prone to stalling had an enrichment for corresponding hydrophobic N-termini that had an association with Hsp70, suggesting the involvement of ribosome associated chaperones in the elongation inhibition mechanism. Overexpression of Hsp70 was shown to prevent the elongation inhibition effect, suggesting that the elongation pausing effect has a cytoprotective function that prevents protein misfolding reducing the burden on the cellular chaperone machinery. Liu et al. (2013) used profiling along with the amino acid analogue AZC, which competes with proline during amino acid incorporation, to induce protein misfolding. They found that the proteotoxic stress triggered ribosomal pausing ~50 codons downstream from start sites in human cells, and hypothesized that the arrest is similarly related to the sequestration of chaperone molecules by misfolded proteins.

The results from some of the aforementioned studies were called into question by a more recent report that examined the role of cycloheximide on ribosome profiling data (Gerashchenko and Gladyshev, 2014). This study produced multiple datasets using seven different concentrations of cycloheximide treatment in either oxidatively stressed or heat shocked yeast cells. This analysis found that cycloheximide treatment at a concentration of 100 µg/ml does not immediately induce complete inhibition of elongation. Rather, the intra-cellular concentration of the drug gradually increases as passive diffusion occurs, which means that after treatment, some initiating ribosomes continue elongating until they encounter the inhibitor molecule, which results in an excess of RPFs over the first few codons of an ORF. The authors of this report detail how the use of cycloheximide can distort measured ORF profiles, and how these artefacts, not oxidative stress nor heat shock, lead to the appearance of inhibition of elongation and a large RPF ‘ramp’ at the beginning of ORFs. Gerashchenko and colleagues propose avoiding use cycloheximide if possible or utilizing a sufficiently high concentration of

the molecule such that the vast majority of ribosomes will be bound to the inhibitor almost immediately after application (as opposed to after roughly two minutes, as observed in the 100 μg/ml experiment). The Gerashchenko study was conducted only using a yeast cells, so further work is needed to evaluate the effects in mammalian cells. Furthermore, the increased density of RPFs at the beginning of yeast ORFs was also observed in another study which specifically did not use any elongation inhibitors (Weinberg et al., 2016).

Several groups have tried to identify transcript features that correlate with local variations in ribosome occupancy in order to quantify factors controlling the rate of translation elongation (Ingolia, 2016). Different algorithms offer differing conclusions about the impact of codon and amino acid usage on translation speed, which may have been biased by the tendency of cycloheximide treatment to redistribute ribosomes slightly from their *in vivo* positions (Hussmann et al., 2015). Ribosome profiling has also been used to show that macrolides, a class of antibiotics, do not primarily block the movement of nascent peptides from the exit tunnel (as was previously thought), but rather, selectively affect the ability of the ribosome to catalyse peptide bonds between certain amino acids (Kannan et al., 2012; Davis et al., 2014; Kannan, 2014). These studies also showed that the antibiotics do not inhibit translation of all bacterial transcripts.

RiboSeq and RNASeq have also been employed to investigate the effects of miRNAs on protein synthesis. An initial investigation found miRNA-induced downregulation of transcript abundance accounts for most (>84%) of decreased protein production at later time points (Guo et al., 2010). Bazzini et al. (2012) qualified this interpretation by showing that at early timepoints in zebrafish embryos, miR-430 first induces translational repression by reducing the rate of initiation and then induces mRNA decay through deadenylation at later timepoints. Stadler et al. (2012) corroborated these interpretations in *C. elegans* embryos, showing that miRNA-targeted, down-regulated mRNAs showed no evidence for premature ribosomal drop-off or long-term ribosomal pausing. Proximity-specific ribosome profiling, which selects biotinylated ribosomes tagged against components of specific sub-cellular compartments, has been used to probe translocation into mitochondria and into the ER (Jan et al., 2014; Williams et al., 2014). A long standing question was resolved when it was shown that mitochondrial inner membrane proteins were co-translationally targeted rather than post-translationally translocated. The recently developed translation complex profile sequencing (TCP-seq) adds a crosslinking step (with formaldehyde) to the RiboSeq protocol, which allows fixing of partial ribosome complexes (Archer et al., 2016). This approach has allowed investigation of ribosome scanning by pre-initiation complexes, revealing that ‘open-

conformation' 5′-UTR scanning small subunits protect ~75 nucleotides and showing that the 40S subunit 'lingers' for short periods of time at stop codons after the 60S subunit has dissociated. The dynamic rearrangements in pre-initiation complexes at start codons was shown to protect 19, 29 and 39 nt fragments.

1.2.6 Identification of translation pause sites

As mentioned previously, nascent peptides can induce pauses in translation, which may have implications for protein targeting and folding (Wolin and Walter, 1988; Kramer et al., 2009; Zhang et al., 2015). 'Arrest peptides' can mediate responses to various environmental cues or facilitate secretory pathways (Lakkaraju et al., 2008; Ito and Chiba, 2013). Analyses of RiboSeq data have shown that segments of positive residues (Dana and Tuller, 2012; Charneski and Hurst, 2013) or prolines (Woolstenhulme et al., 2013) can slow the translational apparatus. One profiling study observed significant levels of translational pausing in bacteria at Shine-Dalgarno like sequences in ORFs (which hybridized with the 16S rRNA of ribosomes), though this has been shown to be an artefact due to gel excision of longer footprints formed by mRNA:rRNA interactions (Li et al., 2012a; O'Connor et al., 2013; Mohammad et al., 2016).

1.2.7 Differential gene expression

Ribosome profiling has allowed for the deconvolution of translational and transcriptional effectors of gene expression, as well as for the characterization of cellular responses that are primarily mediated through translational means. Recent analyses of profiling data have suggested that in the majority of cases, protein synthesis rates are tightly correlated with transcript abundances and that translational effects act as signal amplifiers that increase dynamic range (Csárdi et al., 2015; Weinberg et al., 2016). These studies have suggested that translational efficiency and transcript abundance function in a cooperative model of signal transduction. However, notable exceptions exist in these models, particularly in regards to stress responses that necessitate quick changes in gene expression. Ori et al (2015) integrated ribosome profiling with transcriptomics and proteomics in the context of liver and brain samples from 3 and 24 month old rats. While the study noted that 658 and 490 transcripts (in brain and liver, respectively) had significant changes in RPF counts due to aging, 96 (brain) and 9 (liver) mRNAs had changes solely at the level of translational efficiency. The brain's

greater propensity for translational changes in gene expression was particularly notable in several ribosomal subunits, which were also found to be more abundant in the proteomics data. The authors noted that integrating the various datasets show subtle changes in synthesis and abundance that were not detected as statistically significant when solely using one metric, corroborating previous studies that have postulated that aging related changes in mammalian homeostasis involve subtle fluctuations in multiple systems rather than massive perturbations.

RiboSeq data has shown that IFN-γ regulates human macrophage metabolism and translation by reducing induction of the mTORC1 kinase, which causes translational suppression of repressors of inflammation such as HES1 or other immune mediators such as PPBP and CD109, without concomitant changes in transcript abundance (Su et al., 2015). Influenza virus (IAV), which is known to suppress synthesis of cellular proteins via a variety of mechanisms including cap-snatching and inhibition of cellular pre-mRNA polyadenylation, was previously hypothesized to utilize mRNAs that were preferentially translated compared to their host counterparts. However, RiboSeq data from IAV-infected cells showed that the extensive translation of viral proteins is the result of viral takeover of the mRNA pool in the cell rather than enhanced superior translation efficiency. The study also found that some cellular transcripts were protected or enhanced in their translation activity after IAV infection, particularly those involved in oxidative phosphorylation and respiration. Multiple previous studies have reported that mRNAs with longer poly(A)-tails exhibited increased translation. However, Subtelny et al. (2014) noted that these studies primarily used oocytes and early embryos, and that ribosome profiling of embryonic and non-embryonic cells show that the coupling between tail length and translation efficiency occurs only in early stage embryos and that this correlation disappears in late stage embryos or mature cells.

1.3 Limitations of Ribosome Profiling

Despite the utility of ribosome profiling, care must be taken not to over-interpret data generated by this methodology (Ingolia, 2016). The use of reversible inhibitors such as cycloheximide allows for slow, concentration-dependent elongation to occur prior to cell lysis and results in localised redistribution of RPFs (Gerashchenko and Gladyshev, 2014; Hussmann et al., 2015). This can significantly confound codon-level analyses, which partially explains the disparate conclusions seen in the multiple attempts to investigate the role of codon optimality, tRNA abundance, RNA structure and amino acid composition on elongation kinetics (Kertesz et al., 2010; Siwiak and Zielenkiewicz, 2010; Tuller et al., 2010a, 2010b, 2011; Qian et al.,

2012; Zur and Tuller, 2012; Wallace et al., 2013; Charneski and Hurst, 2013; Rouskin et al., 2014; Yang et al., 2014). Lareau et al. (2014) and Weinberg et al. (2016) have suggested using inhibitor-free lysis to avoid these artefacts, but this results in a wider distribution of RPF sizes and necessitates this entire range be isolated and sequenced. While inhibitor treatment results in localised perturbations, it is not thought to affect gene-level analyses.

Besides the choice of inhibitors, other variations in RiboSeq library preparation can also affect the resulting sequence data. The concentration and type of nuclease (Ingolia et al., 2012; Dunn et al., 2013; Aeschimann et al., 2015), the monosome purification procedure (Aeschimann et al., 2015) and the rRNA depletion strategy (Aeschimann et al., 2015; Chung et al., 2015; Weinberg et al., 2016) can all affect the number and/or phasing quality of the RPF data. *Drosophila melanogaster* ribosomes for example have several unusual rRNA sequences and structures not seen in mammalian or yeast ribosomes, which makes them highly sensitive to RNase I digestion (Dunn et al., 2013). RiboSeq library construction is sensitive to the same sequence biases inherent in RNASeq or generic Illumina sequencing protocols. Previous work has shown that local base composition of transcripts can produce distorting secondary structures, bias reverse transcription priming and inhibit ligation reactions (Zheng et al., 2011), leading to distorted representation of some sequences (Bullard et al., 2010; Hansen et al., 2010; Artieri and Fraser, 2014). A recently developed framework for modelling RPF densities was used to analyse 30 previously published RiboSeq datasets and found a high level of variation in the sequence determinants of RPF frequencies in the various studies (O'Connor et al., 2016). In the mammalian experiments at least, this was partially attributed to the exact timing of cycloheximide treatments as well as lysis and nuclease treatments. Another meta-analysis of 15 profiling experiments revealed a substantial level of noise in measured ribosomal densities at the nucleotide scale and showed that, even when discarding data from 80% of genes with low expression levels, signals are not highly reproducible between experimental replicates (Diament and Tuller, 2016). In particular, only ~30% of RPF peaks were found to be reproducible between replicates.

Analyses of translational efficiency or differential gene expression using NGS requires normalisation between samples to allow for meaningful comparisons. However, these normalisations can be perturbed during scenarios when a large number of genes undergo simultaneous changes in expression (Lovén et al., 2012). Ribosome profiling count data indicates the fraction of active ribosomes that are translating a gene (Ingolia, 2016), so a gene may appear to be differentially translated if there is a global shift in ribosome availability without any concomitant change in the expression levels of that particular gene (e.g. the same

absolute level of a particular protein is being synthesized but a drop in cell-wide, absolute RPF levels increases the fraction of RPFs originating from that particular gene). Transcripts with low RPF or RNASeq counts cannot be reliably identified as differentially expressed without a high level of sequencing depth or numerous replicates (Oshlack and Wakefield, 2009; Liu et al., 2014; Sims et al., 2014; Conesa et al., 2016). Weinberg et al. (2016) found that previous RiboSeq-based estimations of translational efficiency were biased due to the large variability in reported mRNA count levels (while RPF abundances between replicates correlated almost perfectly). Potential alternatives to some of these challenges would be to use synthetic spike-in standards (Jiang et al., 2011), or to obtain absolute measurements of transcript levels using single-molecule RNA fluorescence *in situ* hybridization (smFISH) or qRT-PCR, though this can be time-consuming and technically challenging to perform (Lyubimova et al., 2013; Skinner et al., 2013).

Ribosome profiling provides a steady-state, aggregated snapshot of translational activity, but is unable to give direct information on translation kinetics (Chekulaeva and Landthaler, 2016). Recently developed methods for imaging single molecule translation *in vivo* can provide complementary insights for ribosome activity on specific transcripts (Morisaki et al., 2016; Wang et al., 2016; Wu et al., 2016; Yan et al., 2016). While proximity-specific ribosome profiling allows for detection of translation activity adjacent to certain membranes, in general, selective ribosome profiling is technically constrained by the lysis and biochemical purification steps. Information on sub-cellular localization of ribosomes, such as in neuron dendrites, is lost during the course of profiling, and can be better observed through use of the aforementioned imaging techniques, at least for specific transcripts (Chekulaeva and Landthaler, 2016).

RPF distributions can be contaminated by footprint-sized RNA fragments from structured non-coding RNA (such as tRNAs), large ribonucleoprotein complexes or rRNA (Brar and Weissman, 2015). Bioinformatic strategies can mitigate this problem, but the presence of these fragments detracts from the sequencing depth available to *bona fide* RPFs (Ingolia et al., 2012, 2014). Mapping of RPFs, given their short size, can be ambiguous in cases of gene families or multiple isoforms (Li et al., 2009a; Brar and Weissman, 2015; Robert and Watson, 2015).

Ribosome profiling analyses require careful interpretation, include proper experimental and *in silico* quality controls, data visualization, appropriate sequencing depth and multiple replicates.

1.4 Computational Analyses

A typical RiboSeq dataset generated on a HiSeq platform contains several gigabytes of useful sequence information per replicate, representing between 10^6 to 10^8 reads. Processing of this information entails several common steps as well as custom downstream analyses that are tailored to the particular experimental aims of the study. This introduction will focus on the alignment of RiboSeq and RNASeq data as well as analyses of differential gene expression; further details of other custom bioinformatic analyses will be discussed later in the thesis.

1.4.1 Alignment of short Next Generation Sequencing reads

The first step of any RNA sequencing analysis is preparation of a reference file for sequence alignment. When a reference assembly exists, this can simply mean downloading the relevant file from a public repository such as RefSeq, Ensemble or UCSC Genome Browser (Pruitt et al., 2014; Rosenbloom et al., 2015; Yates et al., 2016). For less studied model organisms, this may necessitate the use of a *de novo* transcriptome assembly program such as Trinity or Velvet/Oases, both of which use de Brujin graphs for parsimonious contig modelling from short reads (Grabherr et al., 2011; Schulz et al., 2012). Afterwards, the genome or transcriptome assembly must be indexed by a relevant short-read aligner such as bowtie (Langmead, 2010). Other alignments algorithms such as BLAST or BLAT are unsuitable to the task of aligning large numbers of reads against a reference genome due to the many CPU hours needed to do so. Indexing strategies such as spaced seed indexing (used by the Maq alignment algorithm) or the Burrows-Wheeler transform (used by Bowtie) allow for the initial creation of memory-efficient representations of the reference assembly that drastically reduce the compute time for aligning millions of reads (Trapnell and Salzberg, 2009). In the context of RNASeq and RiboSeq, many studies elect to map reads to the genome using a 'splice-aware' mapper such TopHat or STAR (Trapnell et al., 2009; Ingolia et al., 2012; Dobin et al., 2013). The splice alignment problem requires more specialised algorithms than mapping short DNA reads against a genome reference due to the presence of exon junction spanning reads that must be handled differently than reads entirely contained in an exon. Tophat maps reads against a reference genome first using Bowtie and collects all unmapped reads. The algorithm then assembles a consensus of covered genomic regions, predicts *ab initio* splices between areas with detectable read coverage and attempts to align the initially unmapped reads against these candidate splice sequences. In some RNASeq analyses, the splice-junction data can be utilized further by

algorithms such as Cufflinks to quantify different mRNA isoforms of each gene (Trapnell et al., 2010). The isoform quantification program Kallisto skips the alignment procedure by using a 'pseudo-alignment' heuristic that drastically reduces runtime requirements while offering comparable levels of accuracy (on simulated reads at least) (Bray et al., 2016).

RiboSeq data are usually first processed to remove low quality reads and adaptor sequences *in silico* using Cutadapt or Fastx_clipper (Martin, 2011). Reads are then mapped to an rRNA reference index, followed by the host genome. Sequence alignment accuracy is dependent on errors during the read amplification and sequencing steps, as well as the quality of the genome assembly and transcriptome annotation. Reads from low-complexity regions, pseudogenes or repetitive genomic loci are more challenging to map; in some analyses these reads are simply discarded while in other studies, specific approaches have been implemented to handle them (Ingolia et al., 2012; Dunn et al., 2013; Andreev et al., 2015b). The entire RiboSeq processing and alignment pipeline can now be done remotely thanks to deployment of RiboGalaxy and other tools on the Galaxy cloud service (Afgan et al., 2016; Michel et al., 2016)

1.4.2 Differential Gene Expression

One of the primary applications of NGS count data is the characterisation of changes in gene expression due to differences between organisms or experimental conditions. Ribosome profiling adds the additional nuance of allowing the deconvolution of changes in translation and transcript abundance (Figure 1.6). The RNASeq or RiboSeq count levels in the experimental replicates can be modelled as a generalized linear model. Sequencing based metrics of gene expression activity are influenced by variation introduced by statistical, technical and biological factors (Wang et al., 2009; Anders et al., 2013) and involves analyses in thousands of observations (genes or isoforms) with only a few replicates per comparison. It is technically unfeasible to fit error models for each gene, but statistical frameworks have been developed assuming discrete probability distributions (typically Poisson or negative binomial) that model well the variance structure from the sampling pool of RNA sequence data. The biological and technical variation in count data within one set of conditions must be estimated and then compared to the mean change between conditions to determine if the experimental variable did in fact introduce a change in expression levels that was highly unlikely to have occurred due to baseline fluctuations. Unlike in some other types of biological experiments, *P*-values are inappropriate in high-throughput experiments because the metric is only statistically valid when a single score is computed; seemingly low *P*-values may in fact be obtained many

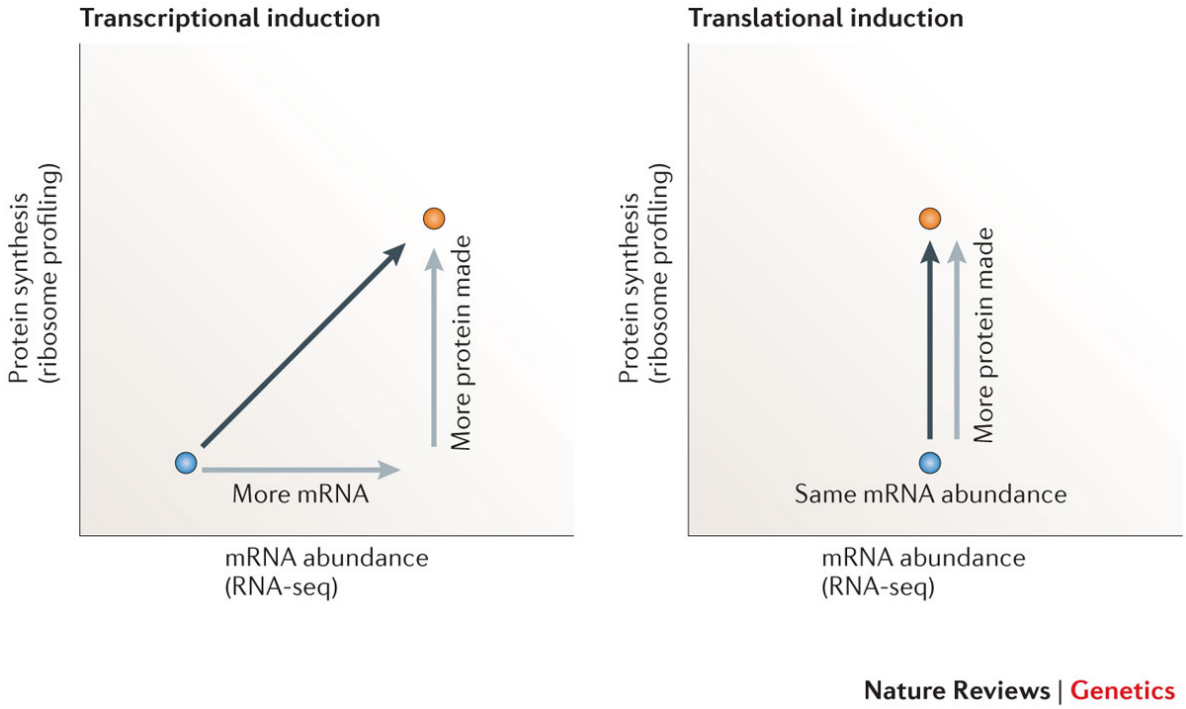


Figure 1.6 Pairing of RiboSeq and RNASeq data allows deconvolution of gene expression mechanisms. Transcriptional induction causes an increase in protein synthesis with concomitant increases in RPF and RNASeq read counts. Translational induction causes an increase in protein synthesis with an increase in RPF read counts but no changes in the RNASeq read counts. Figure from Ingolia, (2016).

times due to the large number of tests performed on an analytic distribution (Noble, 2009). This necessitates the use of multiple testing corrections. One approach uses the Benjamini-Hochberg procedure false discovery rate (FDR) estimation, where *P*-values are sorted in ascending order, and then each observed *P*-value is divided by its percentile rank to get an estimated FDR (such that *P*-values at the bottom of the list correspond to small FDR values) (Benjamini and Hochberg, 1995). This procedure relies on the *P*-values being uniformly distributed under the null hypothesis. A more conservative approach is the Bonferroni adjustment, which controls for the family-wise error rate (the probability of making one or more type I errors when performing multiple hypotheses tests) by adjusting the significance threshold by a factor of $1/n$ for a set of n comparisons (Napierala, 2012). The Benjamini-Hochberg procedure is appropriate when a fixed percentage of false positives in a collection of results are tolerable; the more conservative Bonferroni adjustment is more applicable if focusing on a single result (Noble, 2009).

Increasing the number of replicates per experimental condition increases the statistical power of detecting significant differences between experimental groups. Tightly controlled RNASeq analyses have established that in order to detect significant changes in lowly expressed genes, 12 replicates per condition provide optimal power on a standard HiSeq 2000 (Schurch et al., 2016). Using a smaller number of replicates necessitates discarding lowly expressed genes from analyses or using a more strict false discovery threshold. Similarly rigorous control studies have not been done for ribosome profiling, but the value of multiple replicates is likely to extend to this context as well.

As mentioned previously, comparisons of read count data require normalization to account for differences in sequencing depth and library size. Normalization algorithms that are based on total read counts can be biased when samples have highly heterogeneous transcript distributions (i.e. when some genes are highly expressed or have large changes in expression levels) (Bullard et al., 2010; Hansen et al., 2010; Conesa et al., 2016). The TMM, DESeq, PoissonSeq and UpperQuartile algorithms mitigate this problem by ignoring highly expressed or variable genes (Anders and Huber, 2010; Bullard et al., 2010; Robinson and Oshlack, 2010; Li et al., 2012b; Conesa et al., 2016). Changes in transcript isoform lengths between samples (Trapnell et al., 2013), biases in transcript coverage, non-uniform distribution of cDNA fragment sizes (Roberts et al., 2011) and the GC-content of particular genes (Steijger et al., 2013) can also bias intra-sample comparisons (Conesa et al., 2016).

Numerous statistical frameworks have been developed to allow for analysis of differences in read count data (Table 1.3). Some of the more popular ones include DESeq

Package	Model	RiboSeq Specific?	Reference
EdgeR	NB	No	Robinson et al., (2010)
BaySeq	Bayesian, variable distribution	No	Hardcastle and Kelly, (2010)
Paired BaySeq	Bayesian, variable distribution	Yes	Hardcastle and Kelly, (2010)
Anota	Analysis of Partial Variance	Yes	Larsson et al., (2011)
DESeq	NB	No	Anders and Huber, (2012)
CuffDiff	Beta negative binomial	No	Trapnell et al., (2013)
Babel	NB	Yes	Olshen et al., (2013)
NOISeq	Non-parametric	No	Tarazona et al., (2015)
Limma	Linear regression	No	Ritchie et al., (2015)
RiboDiff	NB	Yes	Zhong et al., (2016)
Xtail	NB	Yes	Xiao et al., (2016)

Table 1.3 Survey of Differential Expression Packages for NGS read count data. Sample of commonly used differential expression packages for inferring differential expression in RNASeq and/or RiboSeq data. 'RiboSeq Specific' denotes whether the given package was designed specifically to handle measurements from ribosome profiling experiments, where there are two types of read count data being normalized and compared (RPF and RNASeq counts). 'Model' describes which distribution and/or statistical approaches are used to model the gene count data and to test for differential expression. 'NB' = Negative Binomial distribution.

(Anders and Huber, 2012), EdgeR (Robinson et al., 2010), Limma (Ritchie et al., 2015) and CuffDiff (Trapnell et al., 2013). The non-parametric NOISeq (Tarazona et al., 2015), which makes minimal assumptions and empirically estimates the null distribution for inferential analysis solely from the read-count data, is particularly suited to studies with few replicates, where estimation of a gamma-Poisson distribution can be noisy (Conesa et al., 2016). The baySeq algorithm (a Bayesian approach based on the negative-binomial distribution) creates multiple models to quantify the differences among experimental groups and estimates the posterior probability of each model for each gene. Several groups have developed RiboSeq specific differential expression packages, including paired BaySeq (Hardcastle and Kelly, 2010), Babel (Olshen et al., 2013), Anota (Larsson et al., 2011), Xtail (Xiao et al., 2016) and RiboDiff (Zhong et al., 2016). Analyses of differential translation efficiency must account for the fact that mRNA abundance estimates also contribute to the denominator in calculations of translation efficiency, so noise in RNASeq data can create negative correlations between transcriptional and translational regulation (Larsson et al., 2011; Katz et al., 2014; Ingolia, 2016).

Multiple studies have used simulated RNASeq reads or used carefully calibrated 'spike-in' additions of known RNA type and quantity to compare the sensitivity and specificity of differential expression tools (Kvam et al., 2012; Nookaew et al., 2012; Robles et al., 2012; Rapaport et al., 2013; Soneson and Delorenzi, 2013; Seqc/Maqc-Iii Consortium, 2014; Seyednasrollah et al., 2013, 2015; Zhang et al., 2016). These studies have demonstrated that the choice of package (or even the particular version thereof) can affect which genes are called as significantly differentially expressed and that no single algorithm is likely to perform favourably for all datasets; hence it is important to evaluate datasets with more than one package (Conesa et al., 2016).

1.5 Aims and Objectives: Survey of translation events via RiboSeq

Ribosome profiling has the capacity to produce a large quantity of data regarding various mechanisms of gene expression *in vivo*. While profiling has already been applied to a number of eukaryotic and prokaryotic organisms as well as several intracellular pathogens, recent work has suggested the need for careful quality controls and data exploration techniques to avoid over-interpretation of RiboSeq results. While the translation apparatus is a broadly conserved component across phylogenetic domains, there are idiosyncrasies in ribosome activity that are specific to certain organisms or even to certain sets of environmental conditions. Comparisons

of translational responses in various organisms or various conditions allow for better characterisation of specific genetic programmes, such as nonsense mediated decay or the unfolded protein response.

This study describes analyses of ribosome profiling data centred largely on the study of RNA virus replication. The studies include the effect of coronavirus infection on murine cell gene expression, the translatome of HIV, the translational kinetics of Dengue virus in both arthropod and human cells and the role of translation in heat and cold shock response in *Arabidopsis thaliana*. During the course of these analyses, several quality-control and visualisation programmes were developed to allow nuanced dissection of RiboSeq data. The results of the analyses corroborate and in some cases, qualify, the interpretations of previous RiboSeq studies as well as those of reports that employed other methodologies.

Chapter Two

Computational Analyses of Retrovirus Translation Data

2 CHAPTER TWO: COMPUTATIONAL ANALYSES OF RETROVIRUS TRANSLATION DATA

2.1 Introduction

One of the aims of this work is to use ribosome profiling to explore the translatomes of various pathogens from a new perspective. While the HIV genome, transcriptome and proteome have been extensively studied, ribosome profiling would allow us to derive a new direct measurement in the context of the viral genome during infection for the HIV programmed ribosomal frameshift efficiency, identify sites of translational pausing and potentially detect non-canonical translation events such as alternative initiation sites or overlapping ORFs. A recent phylogenetic study has proposed the discovery of a new, antisense ORF in the HIV-1 *env* ORF and we sought to verify if this element is translated *in vivo* (Cassan et al., 2016). To provide context for the investigation, this section will provide an introduction to HIV, a well-studied representative of the clinically and veterinarily relevant *Retroviridae*.

2.1.1 Retroviruses

The *Retroviridae* are a family of enveloped RNA viruses that have the distinctive feature of replicating through reverse transcription of their 7-13 kb long RNA genome into dsDNA and subsequent integration of the viral cDNA into the genomes of infected cells (Goff, 2013). Their virions range in size from 80-100 nm in diameter. All retroviruses contain four coding domains with information for virion proteins. The *gag* ORF directs the synthesis of internal virion proteins that form the matrix, capsid and nucleoprotein structures. The *pol* sequence encodes the reverse transcriptase and integrase enzymes. The *pro* ORF encodes the viral protease and *env* contains the sequences for the surface and transmembrane components of the viral envelope protein. Some retroviruses encode additional regulatory proteins which are produced from mRNAs that have been multiply spliced. Retrovirus infection can induce numerous pathologies (including tumorigenesis), and the importance of this class of pathogens has been documented in clinical and veterinary contexts (Table 2.1).

2.1.2 HIV Genome and Life Cycle

Retrovirus example	***Retroviridae* Genus**
Avian leukosis virus	*Alpharetrovirus*
Rous sarcoma virus	*Alpharetrovirus*
Jaagsiekte sheep retrovirus	*Betaretrovirus*
Salmon lymphoma virus	*Gammaretrovirus*
Bovine leukaemia virus	*Deltaretrovirus*
Human T-lymphototropic virus	*Deltaretrovirus*
Human immunodeficiency virus	*Lentivirus*
Equine infectious anemia virus	*Lentivirus*
Human foamy virus	*Spumavirus*

Table 2.1 Clinically and veterinarily important retroviruses. Retroviruses have been shown to cause pathology in clinical and husbandry contexts (Goff, 2013).

HIV belongs to the *Lentivirus* genus whose members are characterised, at least in part, by the possession of cylindrical or conal nucleoid structures as well as several regulatory genes (*tat* and *rev*) which are produced from multiply spliced transcripts (Stoltzfus et al., 2006; Karn et al., 2012). The preferred tropism for HIV is CD4+ cells: macrophages and T-cells, which it infects via the CD4 receptor and CCR5 or CXCR4 co-receptors (Clapham et al., 2001; Coakley et al., 2005). The HIV-1 genome is roughly 9.5 kb in length, depending on the specific strain, and includes a m7G5′ppp5′G_mp cap structure and a poly(A) tail (Figure 2.1). Each end of the genome includes a direct repeat (R), which assists in reverse-transcription as well as integration of the HIV provirus. The HIV-1 genome encodes a total of 15 proteins, most of which are synthesized as components of polyproteins.

Gag ('group specific antigen') encodes the Pr55 polyprotein, which is cleaved by the HIV protease to generate four structural proteins: p17MA matrix, p24CA capsid, p10NC nucleocapsid and p6L1 C-terminal core-envelope link (Goff, 2013). *Env* encodes the Pr160 polyprotein, which is cleaved into surface glycoprotein Gp120/SU and transmembrane glycoprotein Gp41/TM by host furin. *Pol* ('polymerase') encodes a polyprotein that is cleaved to form the reverse-transcriptase (RT) p51, RNase H (RH) p15, integrase (IN) p31, and (PR) p11 (Hill et al., 2005). The polyproteins are cleaved by the *pol*-encoded PR (

Figure 2.2).

The HIV infection cycle begins with the viral gp120 binding to CD4 and CCR5 or CXCR4 co-receptors on the surface of CD4+ T cells and macrophages (
Figure 2.2, Freed and Martin, 2013). The co-receptors play a critical role in HIV entry; individuals expressing a truncated CCR5 mutant allele containing a 32-bp deletion have been shown to be highly resistant to HIV infection and an HIV-positive patient was apparently cured through stem cell transplant from a donor with the truncated allele (Samson et al., 1996; Hütter et al., 2009). After CD4 binding, gp120 undergoes a conformational change which facilitates an interaction with one of the co-receptors. Portions of gp41 embed into the cellular membrane and undergo a conformational change, which facilitates fusion of the viral and cellular membranes (Buzon et al., 2010).

After the viral nucleoprotein complex has entered the cytosol, it begins forming the reverse transcription complex. Cellular $tRNA^{Lys3}$ hybridizes to the primer binding (PB) region of the viral RNA and provides the 3′-OH group needed by reverse transcriptase to initiate minus-strand DNA synthesis (Sarafianos et al., 2009). The reverse transcriptase has an RNase

H domain which degrades the RNA of the resulting RNA:DNA hybrid. The tRNA primer and DNA are repositioned (strand switch) to the 3′ of the RNA genome, and the DNA segment is

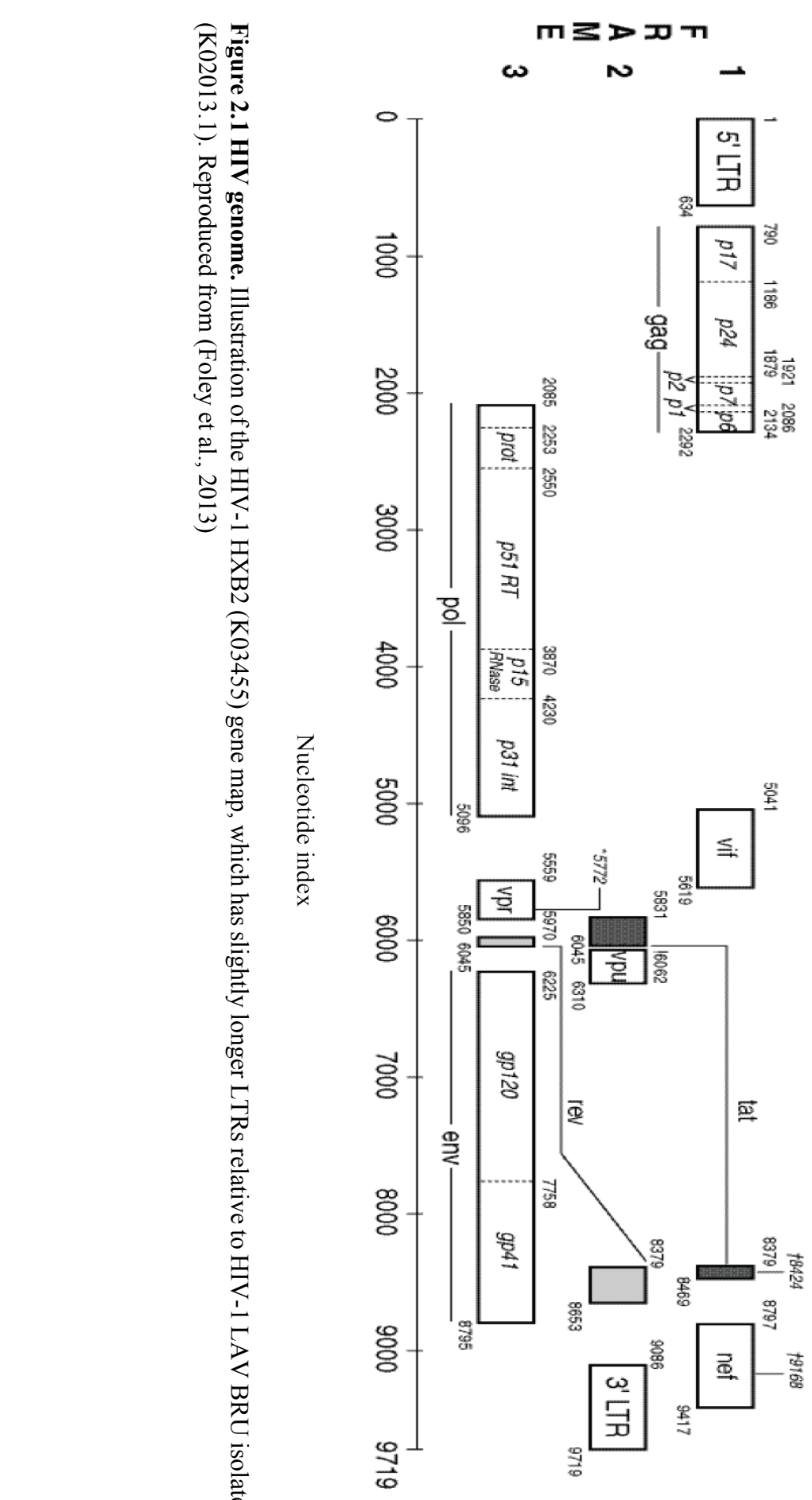


Figure 2.1 HIV genome. Illustration of the HIV-1 HXB2 (K03455) gene map, which has slightly longer LTRs relative to HIV-1 LAV BRU isolate (K02013.1). Reproduced from (Foley et al., 2013)

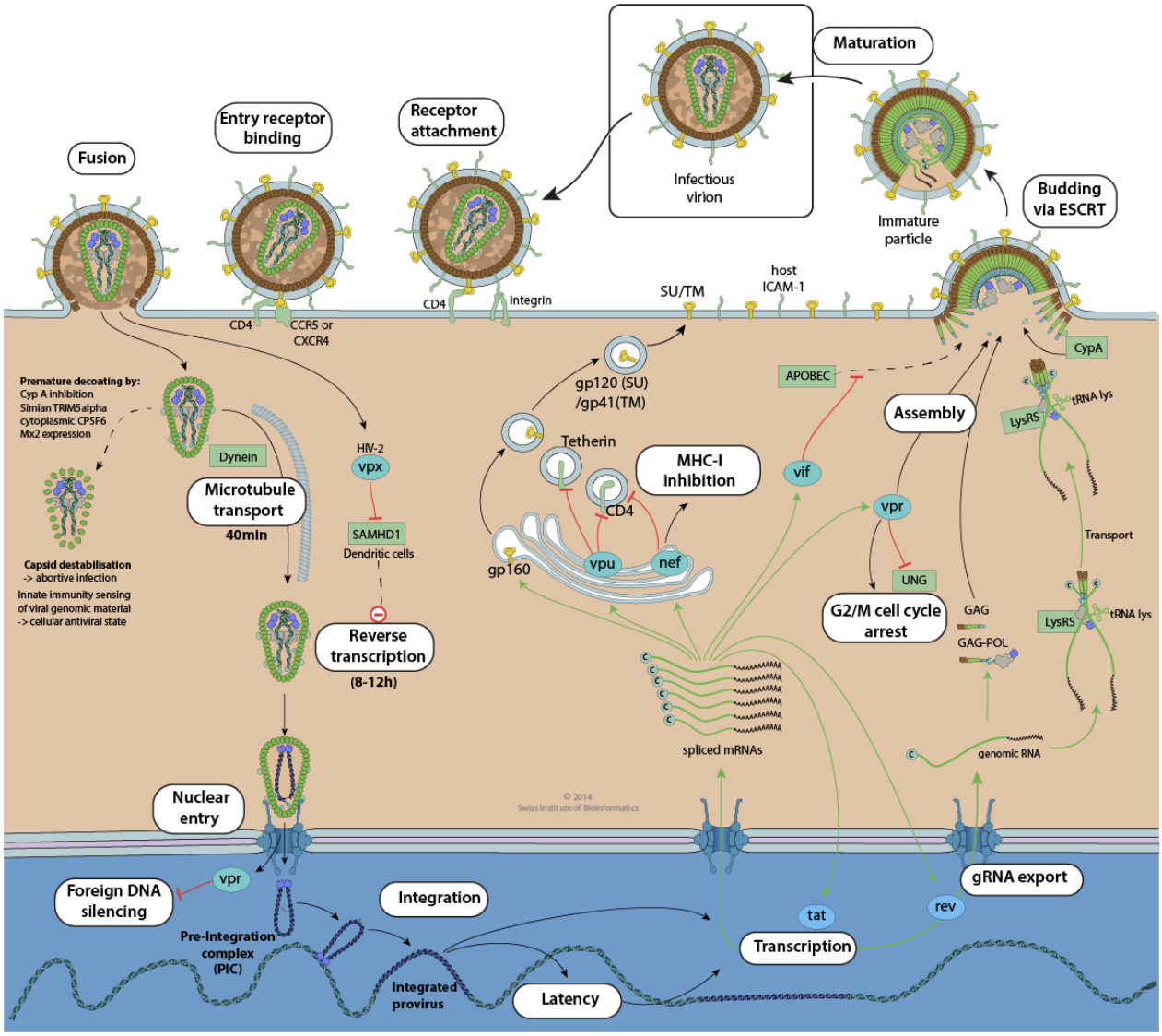


Figure 2.2 HIV replication cycle. Receptor binding proteins on the surface of the virion interact with co-receptors on cell surface. Upon binding, the virion envelope and its associated glycoproteins fuse with the cell membrane. The viral RNA genome is reverse transcribed by a reverse transcriptase carried in the virion. The viral cDNA is transported into the nucleus and the retroviral integrase utilises the LTRs to insert the viral DNA into the host genome. Cellular RNA pol II produces viral transcripts which undergo a sequential process of alternative splicing (with the exception of the full-length transcript from which Gag and Pol are produced) and protein production. After the viral proteins have been synthesized and the full-length virus transcript has been exported from the nucleus, virion assembly and release occur. Figure reproduced from *ViralZone* (Gasteiger et al., 2003).

extended to the 5′ end of the RNA template. RNase H degrades the majority of the original RNA template, except for portions that are used to prime plus-strand DNA synthesis. Positive strand synthesis commences and the tRNA primer is removed, after which the two DNA strands hybridize at the complementary primary binding sites. The reverse transcriptase finishes extending both strands using the newly synthesized DNA segments as the template. The reverse-transcriptase, unlike some cellular polymerases, lacks a proofreading capability, which results in higher levels of copying errors that contribute to high levels of virus sequence variability.

The resulting viral DNA, a copy of the virus coding regions flanked by long terminal repeats (LTRs), assembles into the HIV pre-integration complex (PIC) - a nucleoprotein composite consisting of Vpr, integrase, matrix protein, cellular BANF1 and the DNA, which is thought to be actively transported into the nucleus (Arhel, 2010). Unlike many retroviruses, HIV and other lentiviruses are able to infect non-dividing cells (Yamashita et al., 2004). Integrase cleaves several nucleotides from the termini of the viral cDNA and facilitates cleavage of the cellular DNA, thus allowing the nucleic acids to be joined together. Next, cellular proteins direct gap repair in the resulting DNA hybrid. HIV provirus integration has been shown to occur with a significant preference for actively transcribed genes.

HIV mRNAs are synthesized in a fashion similar to cellular transcripts, including transcription by RNA polymerase II (RNA pol II), 5′ capping, 3′ end cleavage and polyadenylation by the cellular processing apparatus (Karn and Stoltzfus, 2012b; Freed and Martin, 2013). Initially, a subset of transcripts are spliced and exported, after which a group of partially or completely unspliced transcripts are exported via a Rev-dependent mechanism (

Figure 2.3). The fully-spliced transcripts encoding Tat, Rev and Nef are translated in the cytoplasm. Tat functions to increase the efficiency with which the provirus is transcribed by phosphorylating the C-terminal repeat domain of Pol II. Rev has a nuclear localization signal (NLS) that allows it to interact with importin-β and translocate through the nuclear pore, after which it facilitates export of unspliced or partially spliced transcripts (detailed later in the text). The Vpu/Env transcripts are translated by ribosomes on the rough endoplasmic reticulum, while Gag and Gag-Pol are translated by cytoplasmic ribosomes (Freed and Martin, 2013).

The gp160 Env polyprotein is transported through the endoplasmic reticulum and Golgi complex, during which it undergoes cleavage. The gp41 and gp120 glycoproteins embed on the cellular plasma membrane, after which Gag, Gag-Pol and two dimerized copies of the viral genome coalesce. HIV virions bud from the cell surface and are initially non-infectious, immature particles. Subsequently, the viral protease autocatalytically cleaves from the Gag and

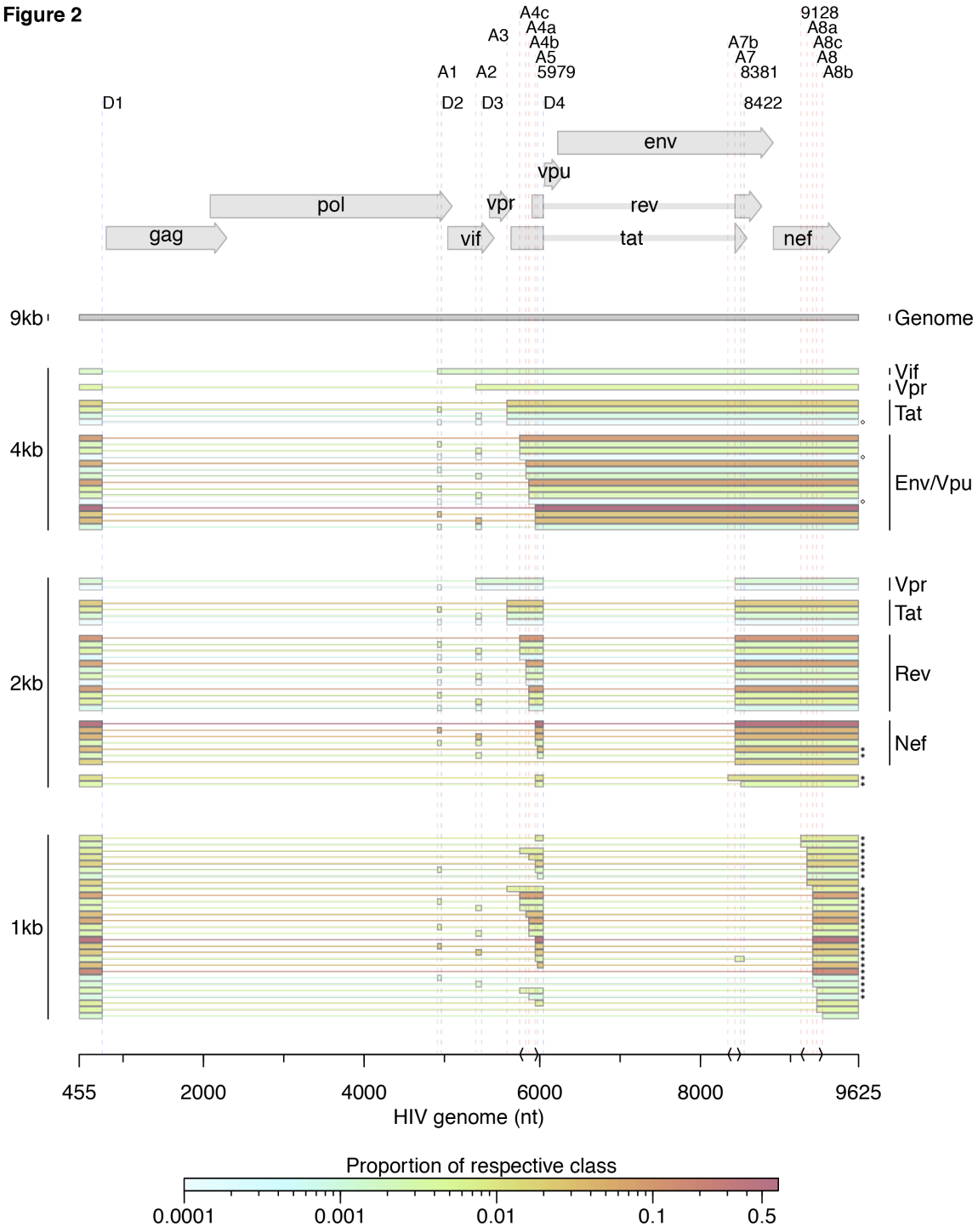


Figure 2.3 Previously identified HIV splice variants. Note that the 5′ leader sequence (which contains the putative uORF upstream of the *gag* AUG) is shared across all HIV splice variants. Figure reproduced and adapted from Karn and Stoltzfus (2012).

Gag-Pol polyproteins and subsequently releases the other functional proteins in a process termed maturation, thereby generating infectious virions.

2.1.3 HIV frameshift signal

The *pol* ORF is expressed as the C-terminal component of a 160 kDa Gag-Pol fusion protein. The extended peptide is synthesized when elongating ribosomes shift into the −1 reading frame upon encountering the ribosomal frameshift element (which is located ~1300 nt downstream of the *gag* AUG). This *cis*-acting motif consists of a heptameric nucleotide sequence (U-UUU-UUA), an 8-nucleotide spacer sequence and a downstream stem-loop structure (Figure 2.4). Given that frameshifting is obligatory for HIV *pol* product expression, it is unsurprising that mutating the PRF sequence or using small molecules to modulate or abrogate frameshifting has been shown to interfere with HIV virion assembly (Shehu-Xhilaga et al., 2001; Marcheschi et al., 2009, 2011). Expression of Gag alone yields non-infectious virions lacking viral enzymes (Cen et al., 2004; Dulude et al., 2006) while solely expressing Gag-Pol results in activation of the viral protease and inhibition of virion assembly (Brierley et al., 2006; Kobayashi et al., 2010; Chamanian et al., 2013).

The efficiency of the HIV-1 frameshift signal has been estimated in *in vitro* translation systems and in transfected cells using a variety of reporter constructs, with estimates varying from 5% to 12.9% (Doyon et al., 1998; Girnary et al., 2007; Mouzakis et al., 2013). The variability in frameshifting efficiencies may be due to random noise, differences in models systems and/or due to the measurement techniques employed. Western blots of Gag and Gag-Pol have been hindered by difficulties associated with transferring large proteins onto nitrocellulose membranes as well as due to some observed intra-cellular processing of Gag-Pol (i.e. cleavage into mature proteins) which may bias protein-based quantification of the frameshifting efficiencies. Differences in ionic concentrations or temperature affecting the stability of the secondary structure, differences in ribosome loading and the presence of competing conformations of the RNA depending on flanking sequences could also contribute to experimental variability (Hori et al., 2016). Lastly, it is also theoretically possible that fine-tuning of frameshifting is partially dependent on the presence of specific intra-cellular co-factors such as non-coding RNA or RNA-binding proteins (Belew et al., 2014; Napthine et al., 2016).

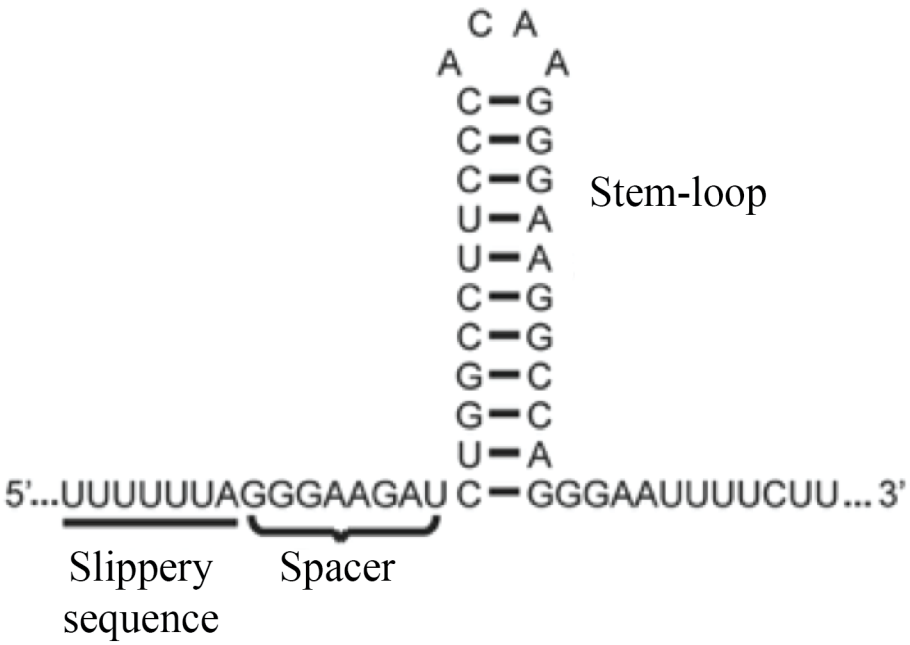


Figure 2.4 HIV frameshift signal. Heptameric slippery site underlined, with folded downstream stem-loop shown. Figure adapted from (Mouzakis et al., 2013).

2.1.4 HIV splicing and Rev Response Element

HIV utilizes an alternative RNA splicing mechanism that allows it to expand the coding potential of its genome in the face of the constraints imposed by the use of a single start site and the dimensions of its virion (Stoltzfus et al., 2006; Karn et al., 2012b). Gag and Pol are translated from a 9.2 kb unspliced genomic transcript; Env, Vif, Vpr, and Vpu are synthesized from singly spliced, 4.5 kb transcripts and Tat, Rev and Nef are expressed from multiply spliced, 2 kb transcripts (

Figure 2.3). Early in the HIV-1 replication cycle, only the fully spliced 2 kb transcript are exported to the cytoplasm for expression of the Tat, Rev and Nef proteins, as the longer transcripts contain intron-like sequences that prevent their export in a fashion similar to intron-containing cellular pre-mRNAs (Taniguchi et al., 2014). The Rev protein translocates to the nucleus due to its nuclear localization sequence and interacts in a 6:1 stoichiometry with the Rev response element (RRE), present on the 4.5 and 9.2 kb transcripts, and cellular Crm1/RanGTP, thus inducing nuclear export of the 4.5 and 9.2 kb transcripts (Daugherty et al., 2010; Fernandes et al., 2012; Sloan et al., 2013). The RRE is a ~350 nucleotide structure consisting of four stem-loops that is found in the *env* CDS of the longer HIV transcripts (

Figure 2.5). NMR studies suggest that Rev binds with high affinity to the IIB stem loop and with secondary affinity to Stem IA (Battiste et al., 1996).

2.2 Ribosome Profiling and RNASeq of HIV-Infected Cells - Data Quality

To analyse HIV RNA and protein synthesis by RNASeq and ribosome profiling, we infected 10^7 SupT1 T-cells (suspended in 25-ml flasks) with the HIV-1 LAV Bru isolate in Category 3 containment (as well as preparing uninfected controls). Ribosome profiling and library preparation was done as described in Ingolia et al., (2012). After two weeks, cells were treated for two minutes with 100 μg/ml cycloheximide (CHX) alone, or 2 μg/ml harringtonine (HAR) for three minutes followed by CHX for two minutes. After drug treatment, cells were pelleted in a Falcon tube for three minutes at 1200 rpm, resuspended in a small volume of the supernatant and transferred to a microfuge tube. Cells were pelleted again, the supernatant removed and the cells resuspended in 400 μl of polysome/lysis buffer (20 mM Tris-Cl pH 7.5, 150 mM NaCl, 5 mM $MgCl_2$). Upon resuspension, samples were triturated by pipetting up and down several times with a narrow plastic gel loading tip. Samples were micro-centrifuged at

4°C for 20 minutes, the supernatant transferred to a Nunc cryo tube, frozen in liquid N_2 and prepared for sequencing as described in section 1.3, including DSN and Ribo-Zero

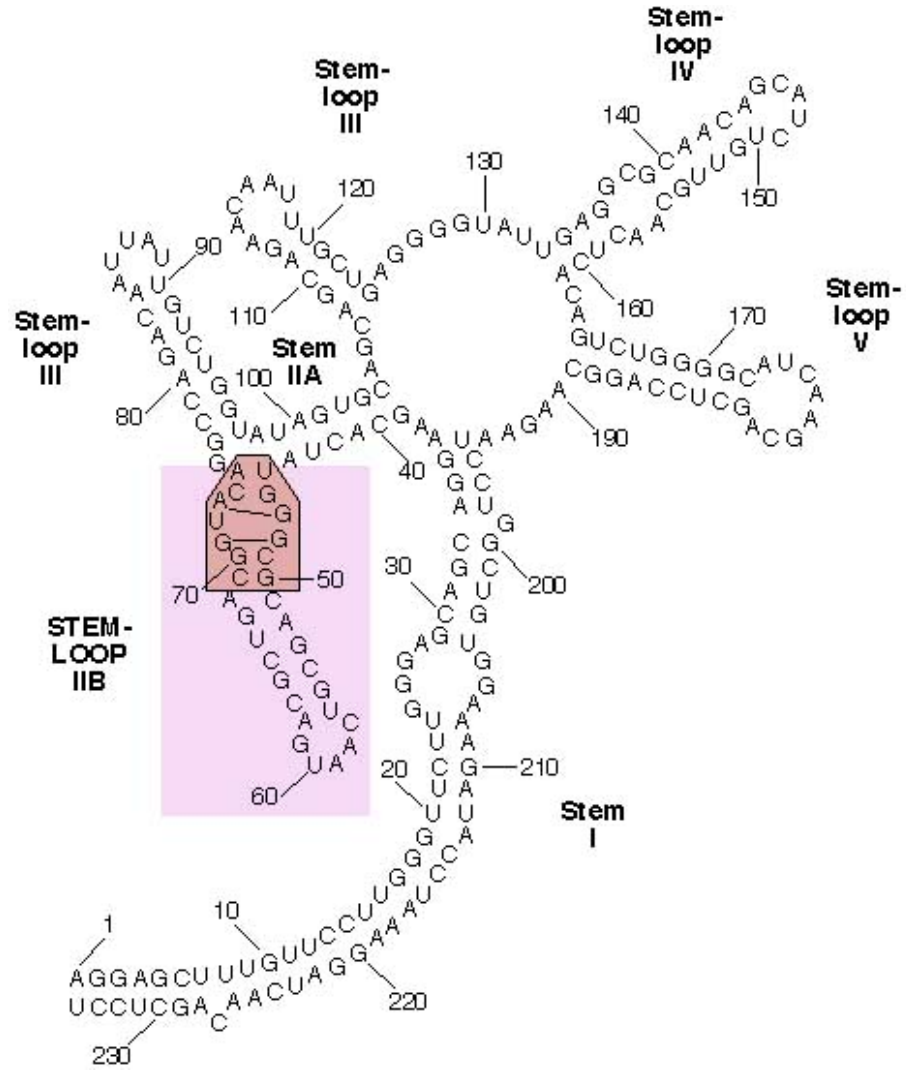


Figure 2.5 Structure of HIV RRE. The RRE consists of several stems, of which at least two interact with the Rev protein during export of various splice variants from the nucleus. Reproduced and adapted from Heguy, (1997). The RRE occurs at positions 7362 to 7595 in the HIV-1 LAV BRU isolate RNA; nt 1 in the diagram corresponds to nt 7362 in the genome.

treatments for rRNA removal in RiboSeq and RNASeq libraries, respectively (Chung et al., 2015). All wet-lab protocols (infections and library preparations) were carried out by Professor Andrew Lever (University of Cambridge, Dept. of Medicine) and Dr Nerea Irigoyen (University of Cambridge, Dept. of Pathology).

Subsequently, RiboSeq (CHX), RiboSeq (HAR) and RNAseq (CHX only) libraries were multiplexed and deep sequenced using an Illumina HiSeq 2000 platform (Beijing Genomics Institute). Reads were trimmed using Fastx-Clipper and mapped to host (UCSC assembly) and virus sequences (NCBI GenBank Accession # K02013.1) (

Figure 2.6). To check that the viral reads corresponded to the HIV reference strain we intended to use, all unmapped and vRNA-mapping reads were assembled using Trinity and queried against the K02013.1 sequence using BLASTn (Grabherr et al., 2011). No insertion or deletions were detected in the assembled sequence. Four substitutions in the virus sequences were detected after assembling viral RNASeq reads using Trinity (Grabherr et al., 2011), but all of these substitutions were synonymous.

The composition of each library is summarized in

Figure 2.7. Both rRNA depletion strategies appear to have successfully reduced rRNA levels in the final libraries. Less than 1% of reads were derived from vRNA, indicating a relatively low level of viral gene expression, despite clear syncytial formation in the T-cell cultures.

2.2.1 Read length distributions

Initially it was decided to perform data quality checks, especially in light of the very low viral expression levels. A comparison of read length distributions of host-derived RNA and virus-derived RNA allows detection of cross-contamination between RiboSeq and RNASeq libraries or between infected and uninfected replicates. In the RNASeq libraries, the reads are expected to have a more uniform length distribution from roughly 27 to 35 nucleotides (determined largely by the gel slice size selection), while RiboSeq libraries are expected to be mostly 28-30 nucleotides in length. Therefore, if RNASeq reads were erroneously mixed with a RiboSeq sample, this would result in a broader distribution of read lengths than in an uncontaminated sample. Our quality-controls revealed very little contamination of this type in both uninfected and infected libraries (Figure 2.8). It should be noted that using short reads for mapping such as those in our libraries presents an additional bio-informatic challenge due to potential non-specific mapping of reads (discussed in Chapter 6). This is not a problem in this

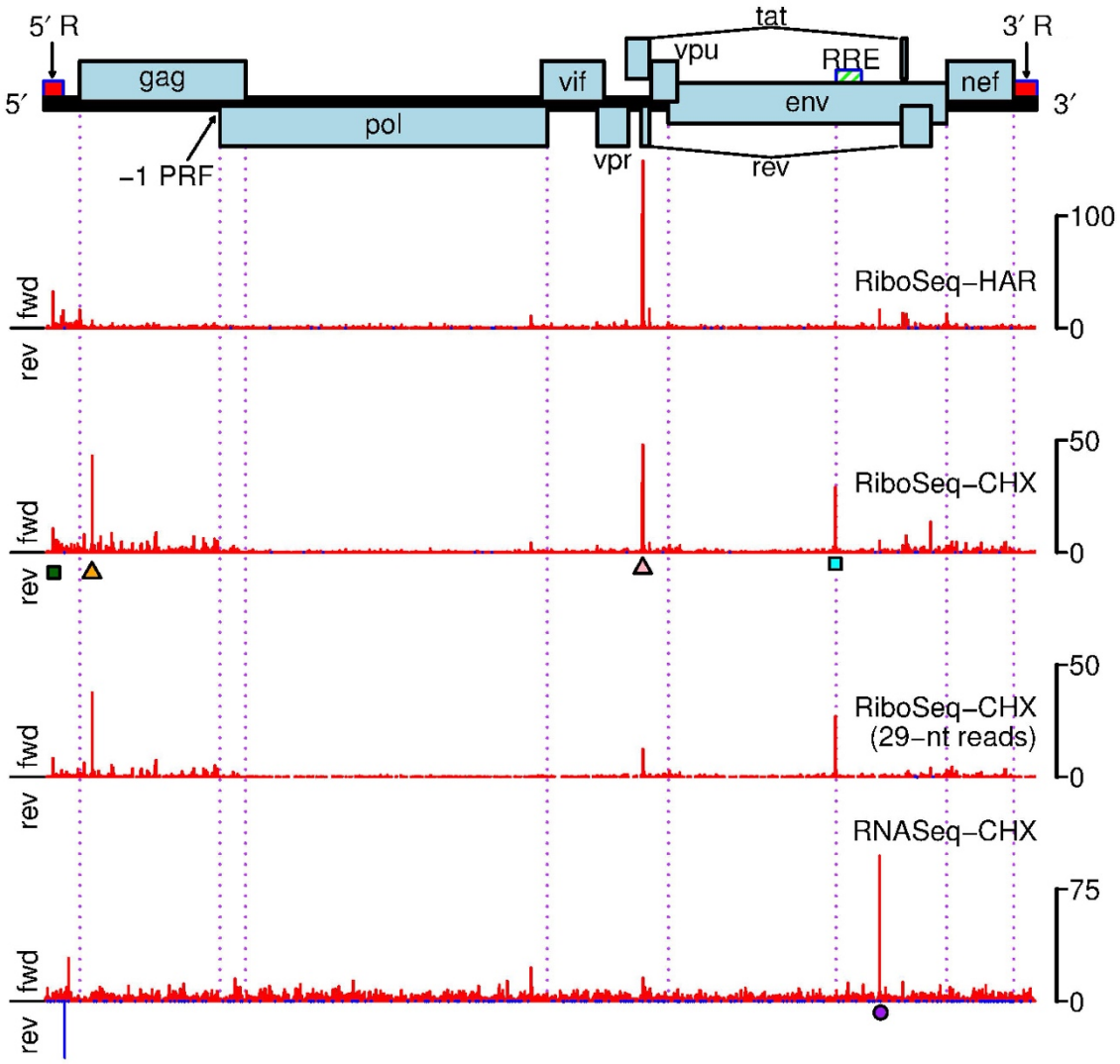
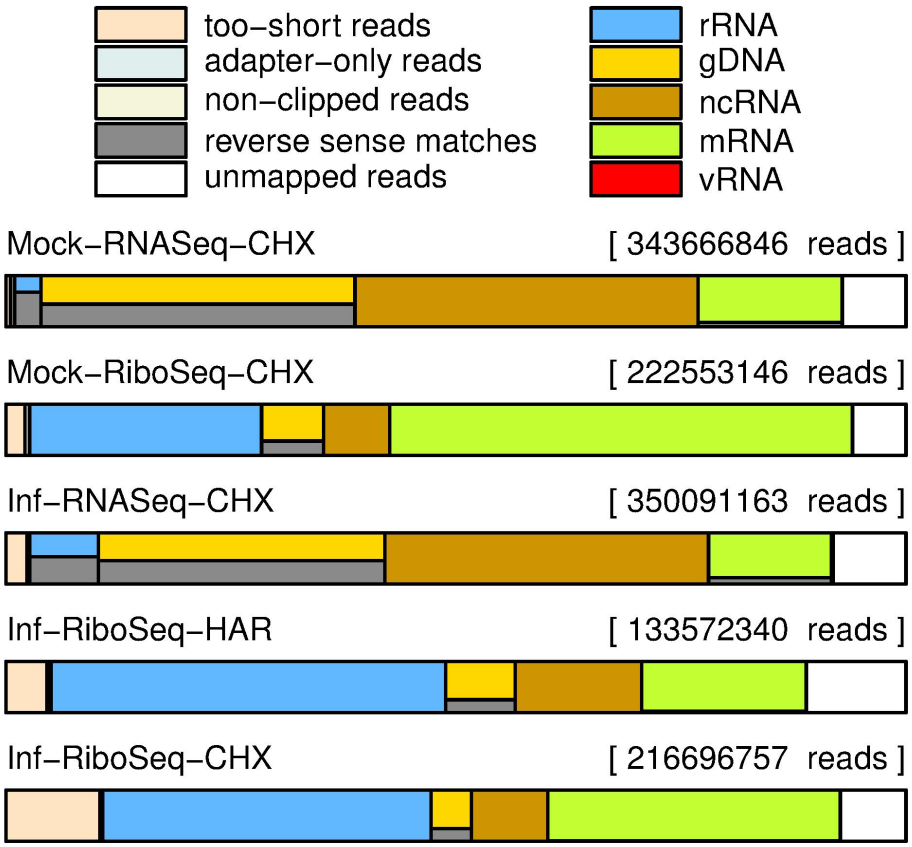


Figure 2.6 Alignments to HIV genome. Map of the 9229-nt HIV LAV genome, with position 1 corresponding to the beginning of the gRNA transcript. Gag and Pol are translated from the unspliced RNA, with Pol being expressed as a transframe fusion product with Gag (i.e. Gag-Pol) via −1 programmed ribosomal frameshifting (−1 PRF). Tat, Rev and Nef are expressed from fully spliced mRNAs, while Env, Vif, Vpr, and Vpu are expressed from partially spliced mRNA. All of the spliced RNAs share a common 5′ leader exon. The Rev Response Element (RRE) is labelled with a green box, and the respective R regions are labelled with red boxes. Histograms depict RiboSeq HAR, RiboSeq CHX and RNASeq densities in reads per million mapped reads (RPM) to the host and viral genomes. For the RiboSeq CHX libraries, both the complete pool of mapped reads as well as a restricted pool of 29-nt reads are shown. Negative-sense reads (including the primer tRNA lysine used by the HIV reverse transcriptase) are show in dark blue below the horizontal axis; the tRNA reads are unlikely to be of viral origin but instead likely represent fragments of host tRNA lysine. Histograms show the positions of the 5′ ends of reads with a +12 nt offset to map (for RPFs) approximate P-site positions. Coloured shapes beneath horizontal bars mark peaks that are described in further detail in the text.



Library	Total Reads x 1000	Trimmed Reads x 1000	rRNA x 1000	(-) vRNA x 1000	tRNA-lysine x 1000	(+) vRNA x 1000	mRNA x 1000	ncRNA x 1000
Mock-RiboSeq-CHX	222,553	216,726	57,378	0.004	0.000	0.163	114,435	16,406
Mock-RNASeq-CHX	343,667	340,277	9,999	0.081	0.077	0.709	55,054	131,027
Inf-RiboSeq-CHX	216,697	193,454	79,112	0.068	0.014	125	70,347	18,398
Inf-RNASeq-CHX	350,091	340,925	26,762	0.375	2	665	47,792	125,909
Inf-RiboSeq-HAR	133,572	126,895	58,597	0.048	0.005	63	24,410	18,775

Figure 2.7 Composition of HIV-1 libraries. The relative levels of RiboSeq and RNASeq reads mapping to various sequence databases. Relatively low levels of rRNA are the result of RiboZero or duplex-specific nuclease depletion for RNASeq and RiboSeq libraries, respectively. **(B)** Table showing raw read count numbers for each library by type. Negative sense vRNA reads are divided into two columns – one containing only the tRNA-lysine reads, and the other showing all other negative sense vRNA reads. In the infected RNASeq library, the majority of negative sense vRNA reads were derived from the tRNA-lysine. 'ncRNA' represents non-rRNA non-coding RNA classes, including tRNA, miRNA, piRNA and snoRNA.

particular analysis as we restricted our downstream analyses to viral reads, for which the 9.2 kb genome is free of non-unique ~30 nt kmers or 30-nt redundancies between the host and viral genome.

2.2.2 Host reads relative to start and stop codons

Figure 2.9 shows the distribution of host mRNA RPF 5′ ends relative to initiation and termination codons, summed over all host mRNAs in each of the RiboSeq libraries. For all samples, a discrete peak in RPF abundance was observed just upstream of the initiation site. As noted previously, the peak is probably largely a result of drug treatment – either HAR which specifically arrests initiating ribosomes, or CHX which arrests elongating ribosomes but allows ribosomes to continue to accumulate at initiation (Ingolia et al., 2011). This peak corresponds to the 5′ ends of RPFs derived from initiating ribosomes with the AUG codon in the ribosomal P-site, and allows calibration of the offset between the RPF 5′ end and RPF P-site position, which, for these libraries, is normally 12 nt. A discrete peak was also observed in the RiboSeq libraries 15 nt upstream of the stop codon, corresponding to ribosomes pausing during termination (with the stop codon in the ribosomal A-site). Averaged over all host mRNAs, very few RPFs were observed in 3′ UTRs while a larger but still low level of RPFs were observed in 5′ UTRs (Figure 2.9). The latter may largely derive from translation of uORFs in various locations and phases with respect to the main ORF of each mRNA (Ingolia et al., 2011).

In the RNASeq libraries there was a peak corresponding to reads with 5′ ends aligning to the stop codon. In both RiboSeq and RNASeq libraries there was a spike at the start codon (in the RiboSeq libraries, this spike occurred four codons downstream of the initiation peak). We suspect that the reads originating from these peaks are primarily the result of ligation biases (and potentially also other biases), as every read mapping to these positions begins with the same 5′ tri-nucleotide sequence (thus accentuating any ligation preferences), while reads derived from the RiboSeq initiation peaks or other sequences will have different 5′ and 3′ ends in different mRNAs.

2.2.3 Read framing distribution

For RPFs derived from non-organellar ribosomes of eukaryotic organisms, mapping of the 5′ end positions to coding sequences (CDSs) characteristically reflects the triplet periodicity

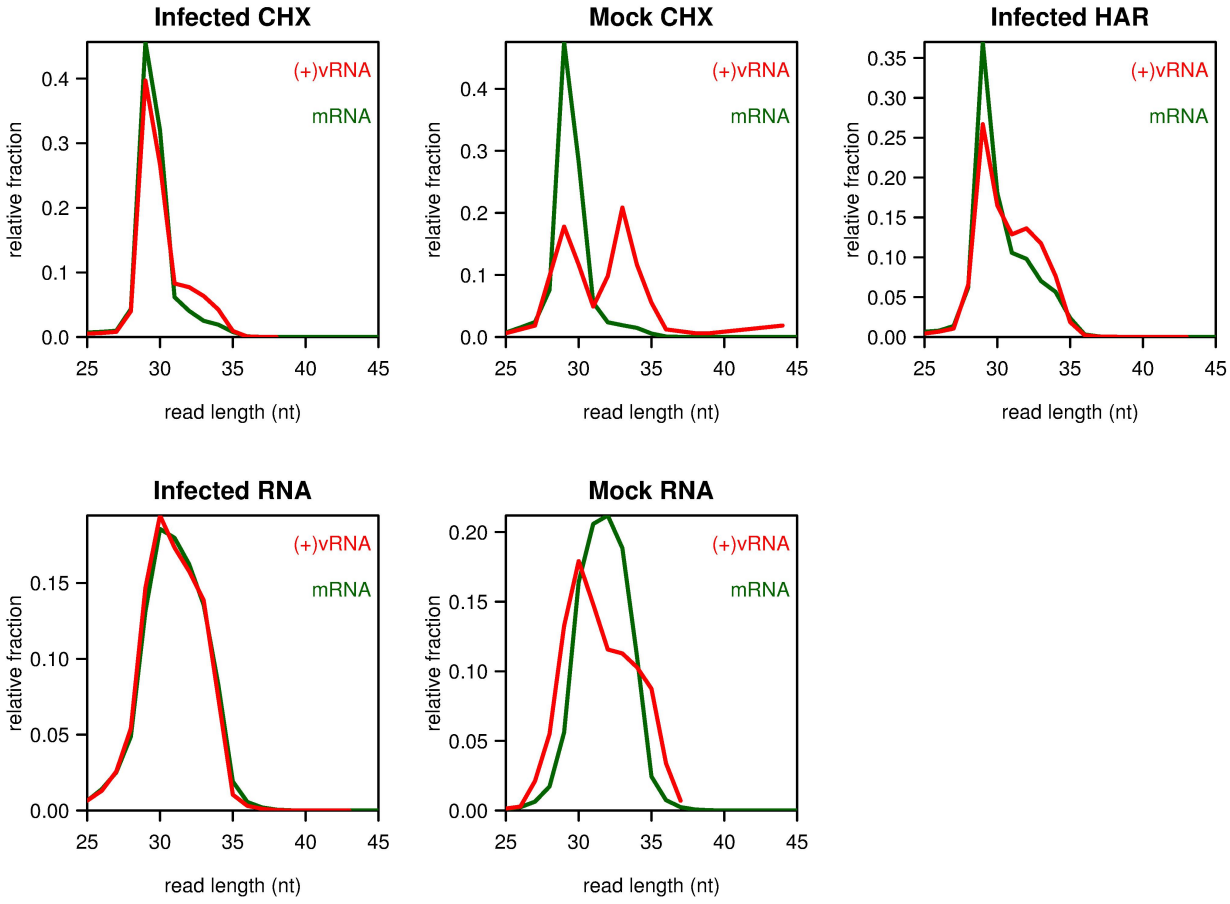


Figure 2.8 HIV library read length distributions. Length distributions for reads mapping to host mRNAs (green) or positive-sense virus RNA (red). Each panel shows the distributions normalized to have equal total sums to facilitate comparison of distribution shapes. RiboSeq libraries tend to have a narrower distribution centred on 30 nucleotides, while RNASeq libraries have broader read length distributions spanning 27-35 nucleotides in length. Mock libraries do appear to have some vRNA contamination, though the absolute levels of these reads are extremely low (see '(+) vRNA' column in table in previous figure).

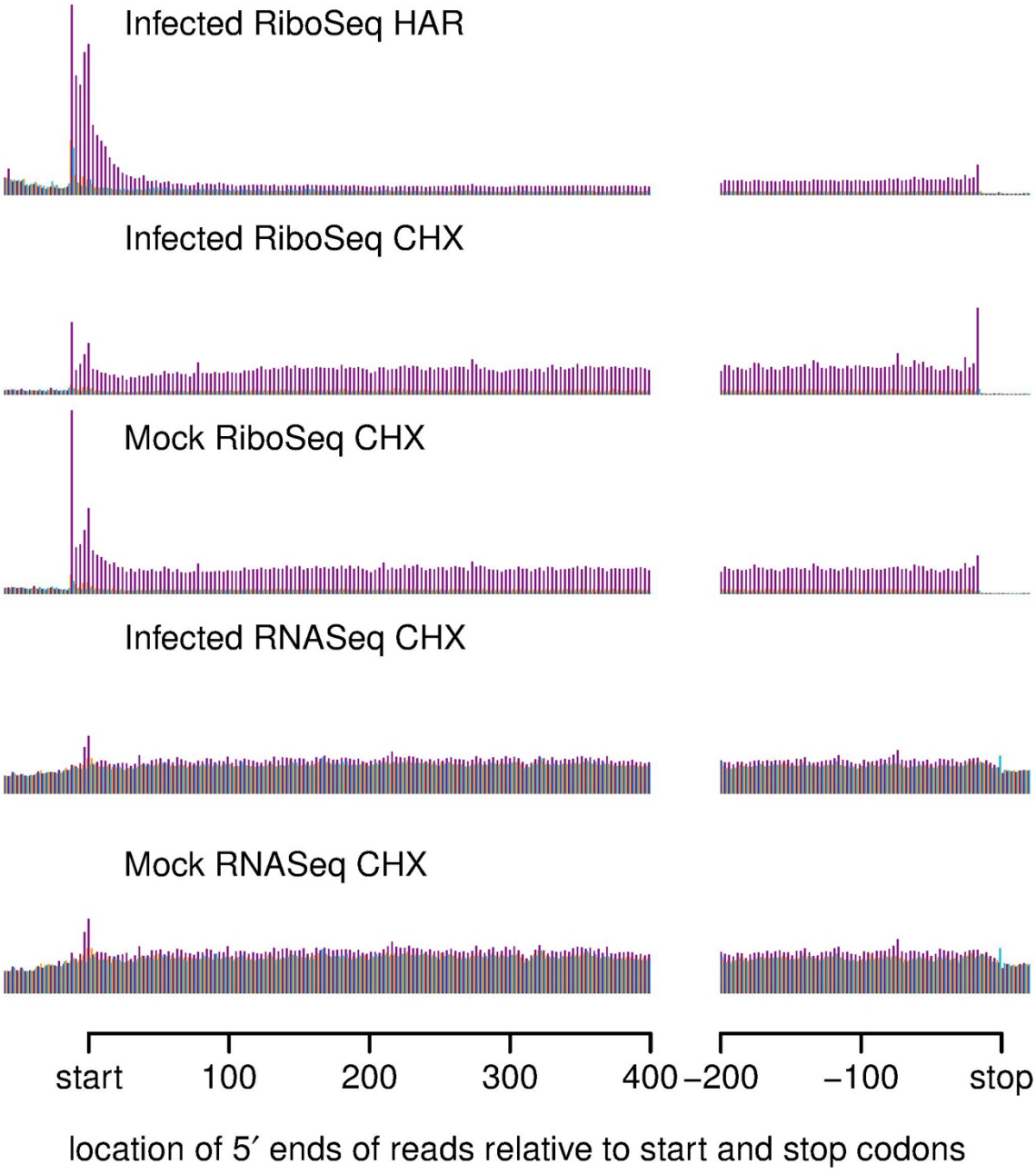


Figure 2.9 Reads mapping to cellular mRNA in HIV libraries. Non-weighted histograms of read 5′ end positions relative to annotated initiation and termination codons summed over all host RefSeq mRNAs for the RNASeq libraries. To account for different library sizes, histograms are normalized by the sum of virus RNA plus total host mRNA for the library, separately for RiboSeq and RNASeq, and then RiboSeq and RNASeq were scaled to have visually similar areas. Reads whose 5′ ends map to the first, second or third positions of codons are indicated in purple, blue or orange, respectively (i.e. a read that maps to the first nucleotide of the AUG initiation codon would be purple.

(herein referred to as "phasing") of translational decoding (Ingolia et al., 2009). Good phasing within datasets provides confidence that RiboSeq libraries are not contaminated with non-RPF RNA, and is beneficial in assigning ORFs with confidence, particularly if such ORFs are very short or overlap. The extent of phasing can vary between protocols and libraries due, presumably, to variation in the efficiency of RNase I (or other nuclease) trimming or other factors. Figure 2.10 shows a histogram of the codon positions to which the 5′ ends of host mRNA reads map for different read lengths. The RiboSeq libraries show excellent phasing with the majority of RPF 5′ ends mapping to the first codon position. Conversely, and as expected, the 5′ ends of RNASeq reads had a nearly uniform distribution between the three possible codon positions. The RiboSeq read length distributions were typically sharply peaked at around 29 nt consistent with other analyses (Ingolia et al., 2011), while those of RNASeq were much broader, consistent with a length distribution set by the size of the gel slice excised during purification of fragmented RNA in the RNASeq protocol (approximately 28-34 nt).

2.2.4 Read coverage of HIV sequences

As an example of the data provided by our experimental strategy,

Figure 2.6 and Figure 2.11 show the density of RiboSeq CHX and RNASeq reads mapping to the virus genome. There is relatively uniform level of RPFs across the entire HIV sequence. The step reduction in RiboSeq density between *gag* and *pol* reveals the proportion of ribosomes that terminate at the *gag* stop codon instead of frameshifting into *pol*. RiboSeq density was also observed in the R region, although not corresponding to known coding regions (see below). RNASeq density was essentially constant across *gag* and *pol*, as well as ORFs from the expected fully and partially spliced mRNAs (with the exception of one position, indicated by purple circle and discussed further below). This suggests that majority of our reads were derived from the full-length transcript, and that the ratio of spliced to unspliced transcripts was low, hence the appearance of relatively uniform coverage. A paucity of reads spanning several known HIV splice sites, such as between positions 5626 and 7972 for the tat/rev encoding transcripts, corroborates this hypothesis.

We observed substantial variability in the RNASeq read depth within a transcript, which we ascribe to biases such as fragmentation bias, PCR bias and ligation bias. Similarly, variability in the RiboSeq data within a CDS may be partly due to nuclease bias, PCR bias and ligation bias but also reflects real variations in ribosome processivity. The current lack of

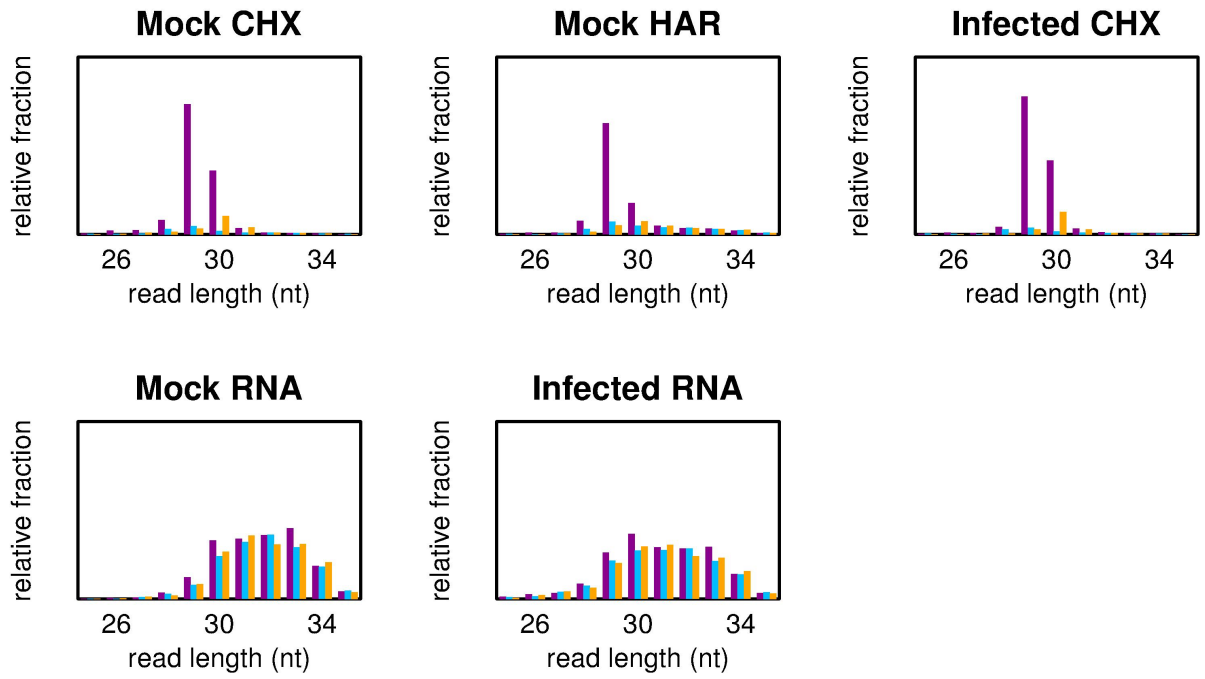


Figure 2.10 RPF Framing Distribution from HIV libraries. Phasing of 5′ ends of reads that map to host mRNA coding regions as a function of read length. Reads whose 5′ ends map to the first, second or third positions of codons are indicated in purple, blue or orange, respectively.

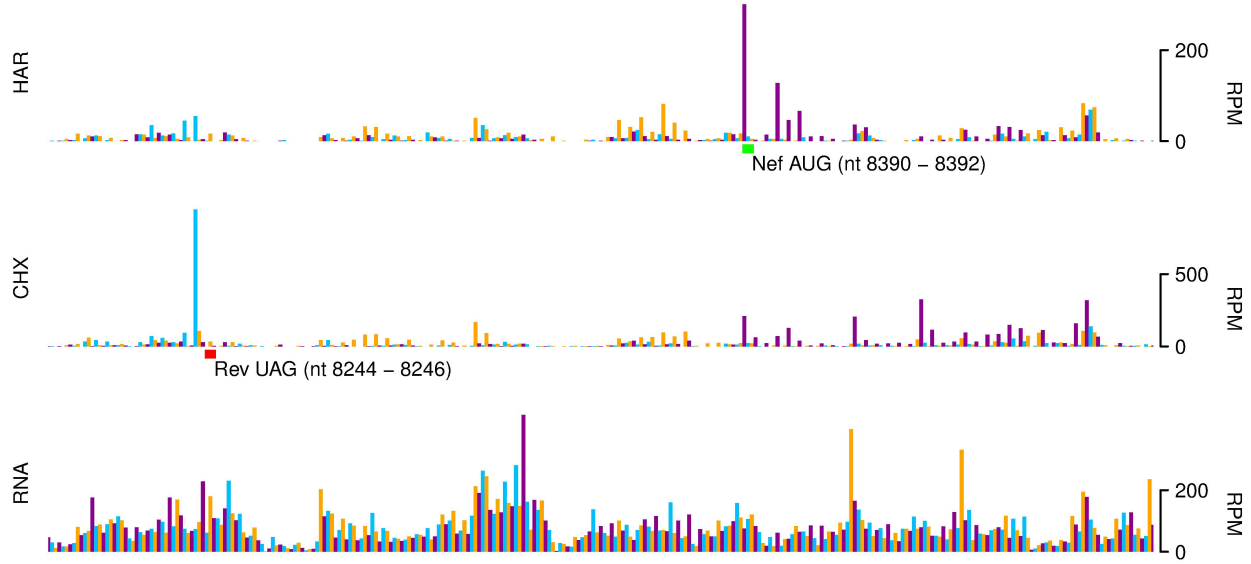


Figure 2.11 Zoom-in plot of *rev* stop codon and *nef* start codon. An example of termination and initiation peaks in HIV. Histograms show the positions of the 5′ ends of reads with a +12 nt offset to map the approximate P-site. Reads whose 5′ ends map to the first, second or third phase relative to position 1 in the HIV LAV genome are indicated in blue, yellow or purple, respectively. The green bar below the HAR library plot indicates the location of the annotated *nef* AUG and the red bar below the CHX library plot indicates the location of the *rev* UAG.

biological repeats prohibits the estimation of the contribution of the sum of these effects in our analyses. Several ribosome pause sites in the RiboSeq data as well as an over-representation of RNASeq reads derived from one sequence were observed and are discussed below.

2.3 Analysis of HIV frameshift signal

2.3.1 Estimation of −1 PRF efficiency

One of the aims of the study was to quantify the efficiency of frameshifting in the context of virus infection. We calculated this value by dividing the RiboSeq density in *pol* by the density in *gag*, with the assumption that the average translation speed is similar between *gag* and *pol* and that the translation is steady state, and found the frameshift efficiency to be 19.8 % (Figure 2.12), This value was based on the mean read densities in the non-overlapping portions of the *gag* and *pol* ORFs, excluding peaks from start and stop codons by only including reads from positions 486 – 1487 for *gag* and 2025 – 4528 for *pol*. Thus the frameshifting efficiency in the context of virus infection is notably higher than previously estimated by indirect methods.

2.3.2 Ribosomes do not pause appreciably at the frameshift site

The relevance of ribosomal pausing to the mechanism of −1 PRF has long been a subject of debate (Farabaugh, 2000; Girnary et al., 2007; Lin et al., 2012). Frameshift signal-associated pauses have been documented in a number of *in vitro* assays (Giedroc et al., 2009; Lopinski et al., 2000; Mouzakis et al., 2013), but there is, as yet, little evidence for a causal relationship, and pausing has not been extensively examined in infected cells. We therefore looked to see whether there was an accumulation of RPFs at the HIV frameshift site, but we failed to see significant pausing here (Figure 2.13). This is consistent with a similar lack of pausing observed in our previous ribosome profiling experiments utilizing murine hepatitis virus (MHV) (Irigoyen et al., 2016) as well as previous toe-printing assays (Kobayashi et al., 2010). An important caveat to consider is that if the stem-loop is closely associated with the ribosome at the mRNA entrance channel, it may be more resistant to RNase I treatment, thus increasing the footprint size of those RPFs. In such circumstances, the gel purification step of RPF isolation would lead to loss of such reads.

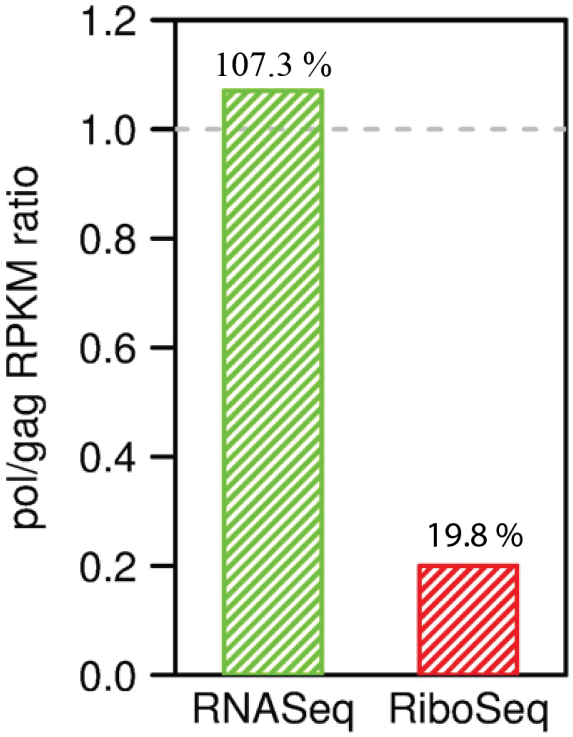


Figure 2.12 Estimate of −1 PRF efficiency. Frameshifting efficiency estimated from the ratio of RiboSeq density in *pol* to that in *gag* (red). For comparison, the same calculation was done for RNASeq (green). *Gag* and *pol* are both present only on the unspliced RNA so the ratio of RNASeq densities in the two ORFs is expected to approximate unity.

2.4 Pause sites in HIV translatome

2.4.1 Site of ribosome pausing in *gag* ORF

Although we failed to identify significant pausing at the frameshift site, there were other sites at which RPFs accumulated to a much higher level than at neighbouring sites. These peaks occur at sequences in various parts of the HIV mRNAs. A substantial pause occurs at position 450 (
Figure 2.14), located in the *gag* ORF, in the portion of the CDS that encodes the highly basic 131 amino acid p17 matrix protein (MA). This pause site occurs nearly 30 codons downstream of the annotated *gag* AUG at position 336. The portion of the Gag peptide encoded by the mRNA sequence immediately upstream of this pause site contains a disproportionate number of positively charged residues (KKKYKLKHIVWASRE; A-site underlined). This is consistent with previous ribosome profiling studies that have documented how the presence of a glutamic acid codon in the A-site is associated with pausing and how specific features of newly synthesized peptides, particularly basic residues can cause ribosome pausing (Ingolia et al., 2011b; Dana and Tuller, 2012; Charneski and Hurst, 2013; Doerfel et al., 2013; Ude et al., 2013; Woolstenhulme et al., 2015). As an alternative possibility, we analysed the RNA sequence downstream of the pause sites for evidence of stable RNA structures that might induce pausing, but no obvious structures were apparent. A further alternative explanation is that these pauses are induced by *trans*-acting factors, e.g. RNA binding proteins, or chaperones of the nascent peptide. One of the documented limitations of ribosome profiling is also the possibility of contaminating footprint-sized fragments, such as from structured non-coding RNAs or large ribonucleoprotein complexes that are isolated in the same sucrose cushion fraction as ribosomes (Ji et al., 2016).

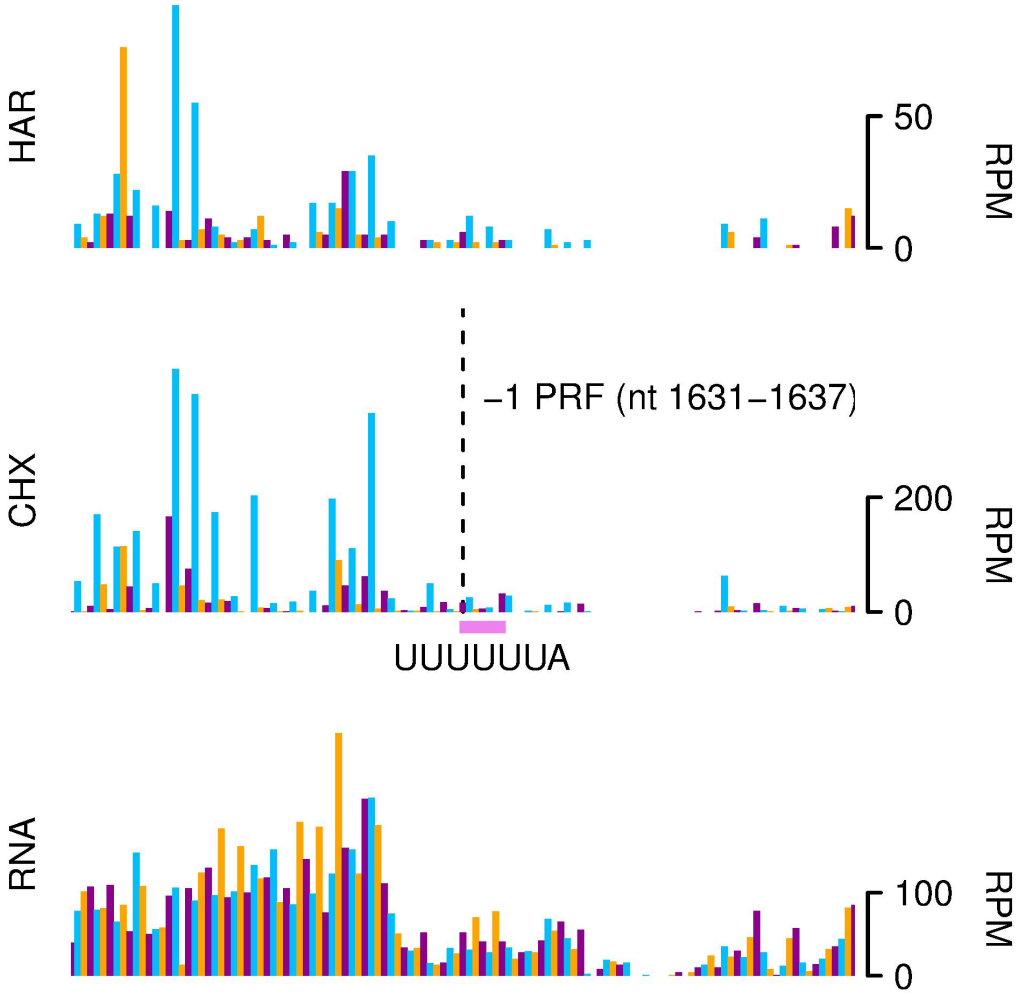


Figure 2.13 Zoom-in plot of positions 1570-1690. No significant pile-up of RPFs was detected at the HIV frameshift site.

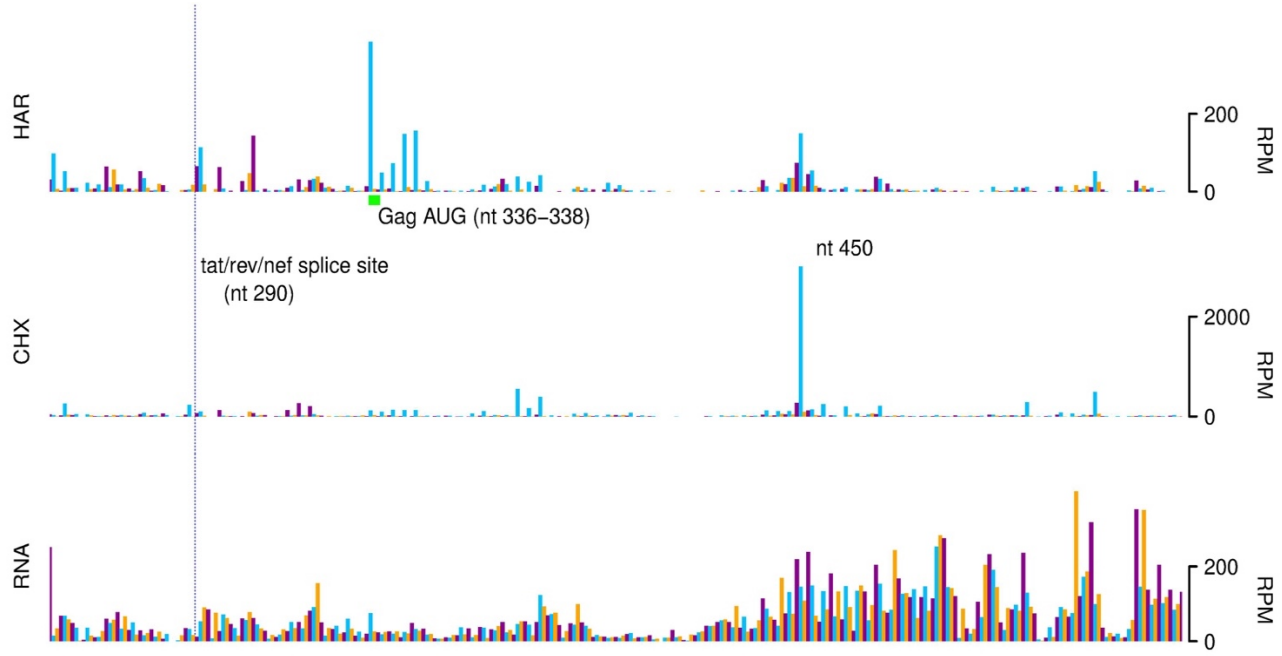


Figure 2.14 Enlarged view of the pause site in the *gag* ORF. The pause occurs when ribosomes have their P-site positioned at coordinate 450. The most abundant RPF read at this position is GUAUGGGCAAGCAGGGAGCUAGAACGAUU (29 nt) and the Gag peptide at this location is KKKYKLKHIVWASRELERF (A-site underlined). This pause site occurs more than 30 codons downstream from the annotated *gag* AUG at position 336-338. The dashed purple line at nt 290 represents the tat/rev/nef splice site.

2.4.2 Site of ribosome pausing in *rev* and *tat* ORFs

A pause site was observed in the *tat*/*rev* ORFs, in both harringtonine and
libraries, at positions 5563 to 5567 (

Figure 2.15). This occurs 12 to 16 nt downstream of the rev AUG. Though the *rev* AUG exhibits an initiation peak in the harringtonine library at position 5551, it is ~ 45 to 70 times smaller than the RPF peaks at positions 5563 to 5567. No AUG start codon occurs in this sequence (in any frame) and no strong predictions were obtained for a downstream RNA structure. The reads mapping to this position mapped uniquely; there was no potential for multi-mapping to other portions of the HIV genome or the host genome. The phasing of the pause RPFs is more consistent with translation of the *tat* ORF rather than the *rev* ORF. The overlapping *tat* ORF at positions 5553-5588 encodes an 11-amino acid sequence with numerous positively charged and proline amino acid residues (GRKKRRQRRRPP; probable A-site underlined). This is consistent with previous work showing that positive charges or prolines in an emerging peptide can significantly slow an elongating ribosome, potentially through interactions with the negatively charged ribosomal exit tunnel (Dana and Tuller, 2012; Charneski and Hurst, 2013; Doerfel et al., 2013; Ude et al., 2013; Woolstenhulme et al., 2015). Previous studies have shown that the GRKKR sequence in the Tat protein functions as a nuclear localization signal, similar to how many other nuclear import signals contain tracks of basic amino acid residues (Ruben et al., 1989; Ragin et al., 2002), so translational pausing may be an unavoidable consequence of having to encode an NLS.

2.4.3 Site of ribosome pausing upstream of Rev response element

We observed a large pause site immediately upstream of the RRE (positions 7362 to 7595) in the *env* ORF at position 7357 (

Figure 2.16), which may be due to the influence of Rev proteins binding the mRNA, the large RNA structure or both. The Rev RNP complex might inhibit ribosome processivity, leading to translational stalling, or the Rev RNP may co-sediment with ribosomes during the RPF isolation protocol leading to contamination. The latter seems less likely as the observed pause is just upstream of the RRE and 98.9% of reads aligning to this locus are 29-30 nt in length.

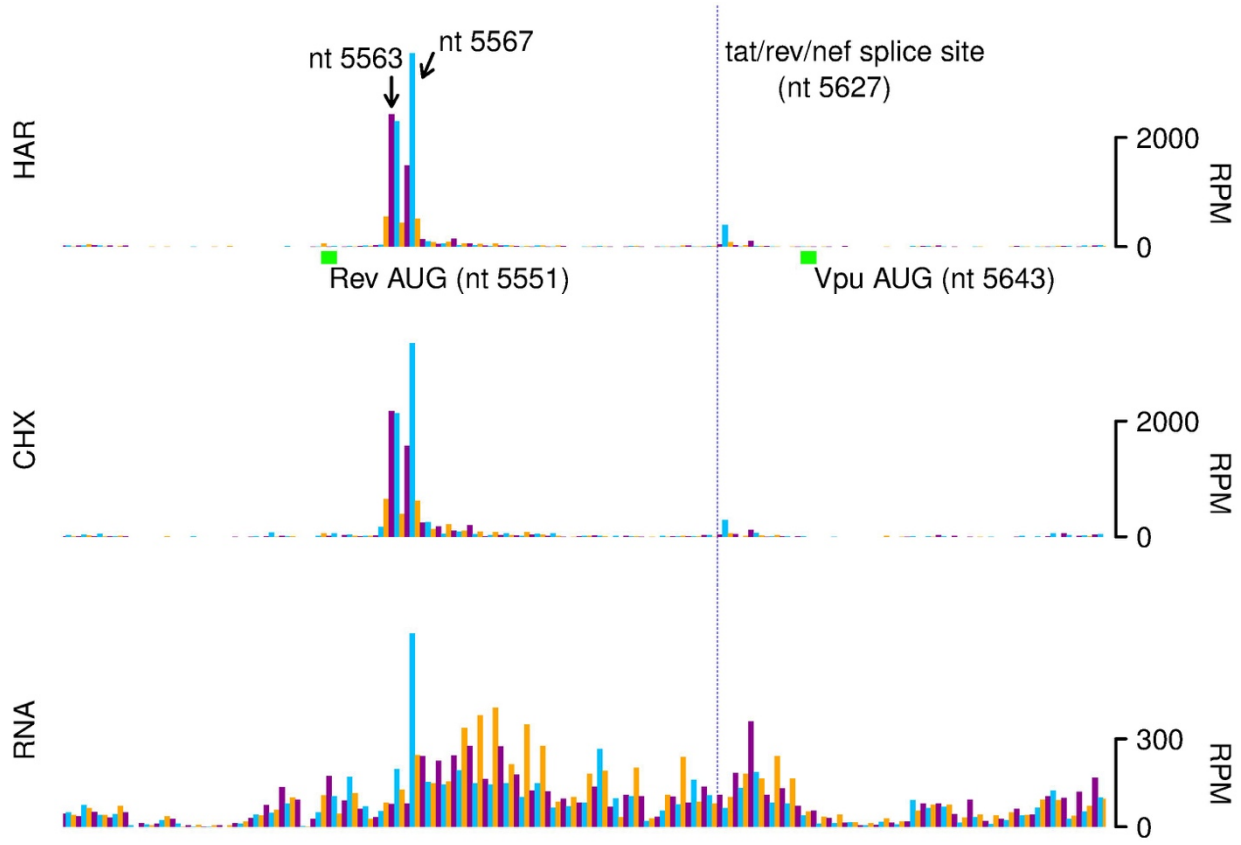


Figure 2.15 Histogram indicating RPF and RNASeq reads mapping to positions 5500-5700 (*rev* and *tat* ORFs) on the HIV genome. Histograms show the positions of the 5′ ends of reads with a +12 nt offset to map the approximate P-site. Green boxes underneath the harringtonine plot indicate the position of the *rev* AUG and the *vpu* AUG. The dashed purple line indicates the location of the exon-intron boundary for the *tat*, *rev* and *nef* spliced transcripts. The harringtonine initiation peak at the *rev* AUG is significantly smaller than the peak observed at positions 5563 to 5567.

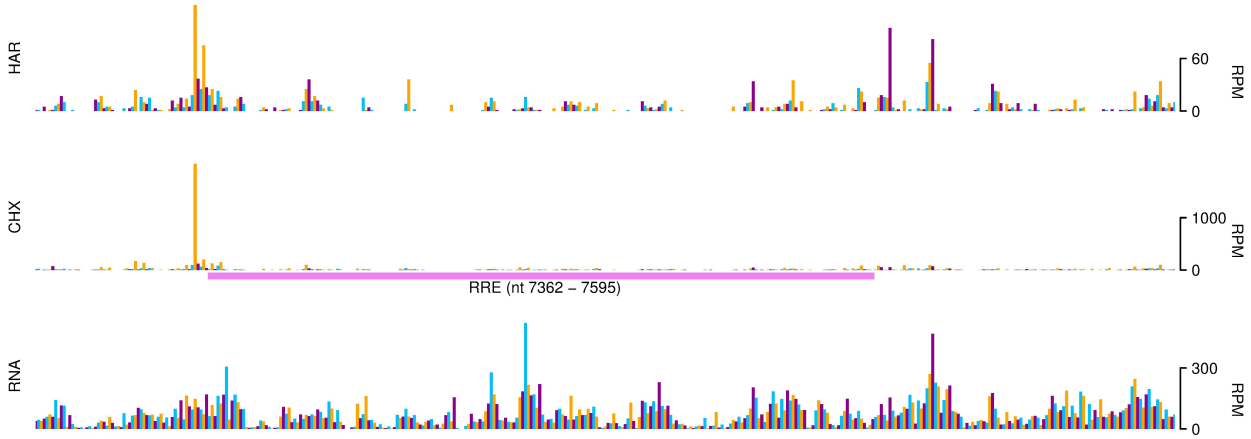


Figure 2.16 Histogram indicating RPF and RNASeq reads mapping to positions 7300-7700 (*env* ORF) on the HIV genome. The violet horizontal bar indicates the location of the RRE at positions 7362 to 7595 (as shown in
Figure 2.5).

2.5 RPFs upstream of gag ORF

Upstream of the *gag* AUG at position 336, we observed a peak that corresponded to a UUG-uORF (

Figure 2.17). Upstream ORFs are present in ~40% of mammalian mRNAs and have been shown generally to cause repression of translation of the downstream (main) ORF (Calvo et al., 2009; Somers et al., 2013), though some have been shown to also encode functional peptides (Menschaert et al., 2013). We observed RPFs mapping specifically to and in-frame with the uORF, as well as an accumulation of reads at the first downstream in-frame stop-codon, confirming that it is translated. It appeared to be translated as efficiently as *gag*, though it has a moderate initiation context (gccUUGa) and UUG is generally regarded as one of the less efficient initiation codons in eukaryotic mRNAs (Ivanov et al., 2011; Asano, 2014). A few examples of UUG initiated ORFs in mammals have been studied, particularly in N-terminal variants of neurotrophin 3 (Ntf3) and transcriptional enhancer factor Tead1 (Stewarta et al., 1996; Ivanov et al., 2011).

As discussed in Chapter 1, after translating a short ORF the mammalian 40S ribosomal subunit may remain associated with the mRNA and resume scanning until encountering another start codon and re-initiating translation (Kochetov et al., 2008; Obermayer and Rajewsky, 2014). Changes in use of a uORF can modulate the translational efficiency of the main downstream ORF. The start codon context in mammalian cells affects the probability of initiation occurring at a given start codon; a purine at the -3 position and a guanine at position +4 relative to the beginning of the start codon are favourable initiation contexts (Kochetov et al., 2008). The distance separating the termination codon of this uORF from the *gag* AUG (~200 nt) makes it unlikely that it inhibits ribosomal access to *gag* (or the spliced mRNA main ORF AUGs). Given that the uORF is entirely upstream of the 5′ splice site of the sub-genomic mRNAs, it would be present in the 5′ UTR of all of the transcripts. Without a robust quantification of the various HIV transcripts, we could not make estimates of how each isoform contributes to uORF density. On any one transcript species it is likely that only a small fraction of ribosomes (on the order of ~10%) initiate on the uORF, but if it is translated on all transcripts, this would lead to RPF density appearing similar to *gag*. We analysed 12 different HIV-1 strains, including 9 different M subtypes and one example of the N, P and O strains to see if the putative uORF was phylogenetically conserved. The sequence was conserved in the B, D, F1 and G subtypes of the M strain but not in the other M subtypes or other strains. Hence the element does not appear to be widely conserved across different HIV strains.

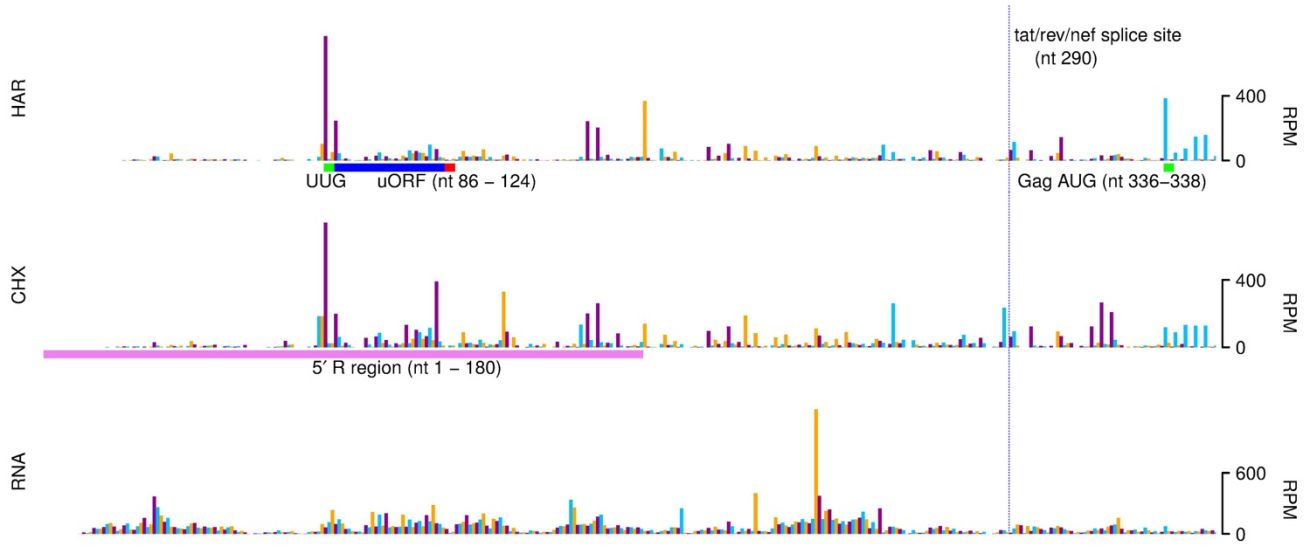


Figure 2.17 Zoom-in plot at positions 1-350. The putative uORF is highlighted in blue, with the UUG start codon underlined in green and the first in-frame stop codon in red. The 5′ R region is underlined with a pink bar, and the *tat*/*rev*/*nef* splice site is indicated with a purple vertical line. Given the position of the uORF upstream of this splice site, it would be contained in all of the HIV splice variants.

2.6 High abundance RNAs

An unexpected over-abundance of RNASeq reads was observed mapping to a unique location within the *env* ORF at position 7755. The reads mapping to this position did not contain significant sequence similarity to other parts of the HIV or human genomes, thus they were uniquely mapped. Previous studies have presented evidence for viral (including retrovirus) encoded miRNAs (Hussain et al., 2012; Kincaid et al., 2012; Klase et al., 2013; Pfeffer et al., 2004; Schopman et al., 2012). The length of the majority of the reads mapping at this position (30 nt) reduces the likelihood that these reads correspond to a mature siRNA or miRNA, although our gel size selection (and library preparation protocol) would exclude sequencing of mature miRNAs regardless. The reads map ~180 nt downstream of the 3′ end of the RRE, making it unlikely that the RRE structure contributes to this particular peak. No similar over-abundance of reads was observed in either the cycloheximide or harringtonine libraries for this sample. The peak may be due to ligation bias but further investigation is needed to ascertain the source and nature of this RNA fragment.

2.7 Estimation of translation efficiency and transcript abundance

We attempted to identify human mRNAs that were up-regulated or down-regulated at the level of transcript abundance or translation efficiency using the R packages DESeq and Xtail (Anders and Huber, 2010; Xiao et al., 2016). However, no statistically significant differences were observed (using a False Discovery Rate (FDR) cut-off of 0.05). This could be attributed to the lack of replicates, which greatly limits the ability to do statistical analyses, but also to the possibility that the virus load at the time of lysate preparation was not sufficiently influencing cellular gene expression. A greater depth of sequencing and additional replicates would be needed to detect HIV-induced changes in host gene expression, if they do occur at significant levels under these conditions.

2.8 Analysis of HIV splicing

Previous RNASeq analyses have been able to estimate transcript abundances for various viruses by calculating the number of RNASeq reads relative to the lengths of their respective transcripts and accounting for the relative expression levels from overlapping, transcript

encoding sequences (Irigoyen et al., 2016). However, given the patterns of alternative splicing in the HIV transcriptome, such an approach is not feasible for unambiguously estimating the abundances of most of the HIV transcripts. Transcript abundances could potentially be calculated from the relative abundances of RNASeq reads spanning each splice junction. Such "chimeric" reads (where the 5′ part maps to the splice donor and the 3′ part maps to the splice acceptor region) were not included in the initial mapping to the virus genome, but were identified subsequently by taking unmapped reads and splitting them into substrings that were compared against different points in the virus genome for matches. Figure 2.18 illustrates the positions of chimeric reads relative to the HIV genome, with the area of each point proportional to the number of reads occurring at a given position. The large number of points on the diagonal axis corresponds to reads with a short (one to two nucleotide) sequence difference relative to the reference, which may be due to library sequencing or preparation artefacts, point mutations in the viral nucleic acids, or differences between the reference genome and our virus strain. Only reads spanning the 5′ donor at position 290 were detected, possibly due to a low abundance of spliced transcripts in the infection (in agreement with the relatively uniform coverage of RNASeq reads across the HIV genome). Thus we could not use this method to determine transcript abundances, though future work with longer RNASeq reads may alleviate this problem.

2.9 Discussion

To our knowledge, this study represents the first high-throughput analysis of the HIV translatome. The results corroborate previous work showing that ribosomes do not accumulate at the wild-type PRF sites in virus-infected cells (Kobayashi et al., 2010; Irigoyen et al., 2016). We have identified several novel sites of ribosome accumulation, some of which are consistent with previous observations/models of stimulus-specific ribosome pausing, such as nascent-peptide induced pausing or RNA secondary structure dependent pausing. We also detected an unusual RNASeq peak that aligns with the *env* ORF at position 7755, which may constitute a hitherto undetected ncRNA. A recently published study used phylogenetic evidence to propose the existence of an antisense ORF embedded in the HIV-1 *env* ORF (Cassan et al., 2016). In our ribosomal profiling data, however, we did not observe any RPFs mapping to this putative ORF, suggesting that it is not translated, at least in the LAV/SupT1 system employed here.

Though our study presented an initial survey of such pause sites, to fully characterise

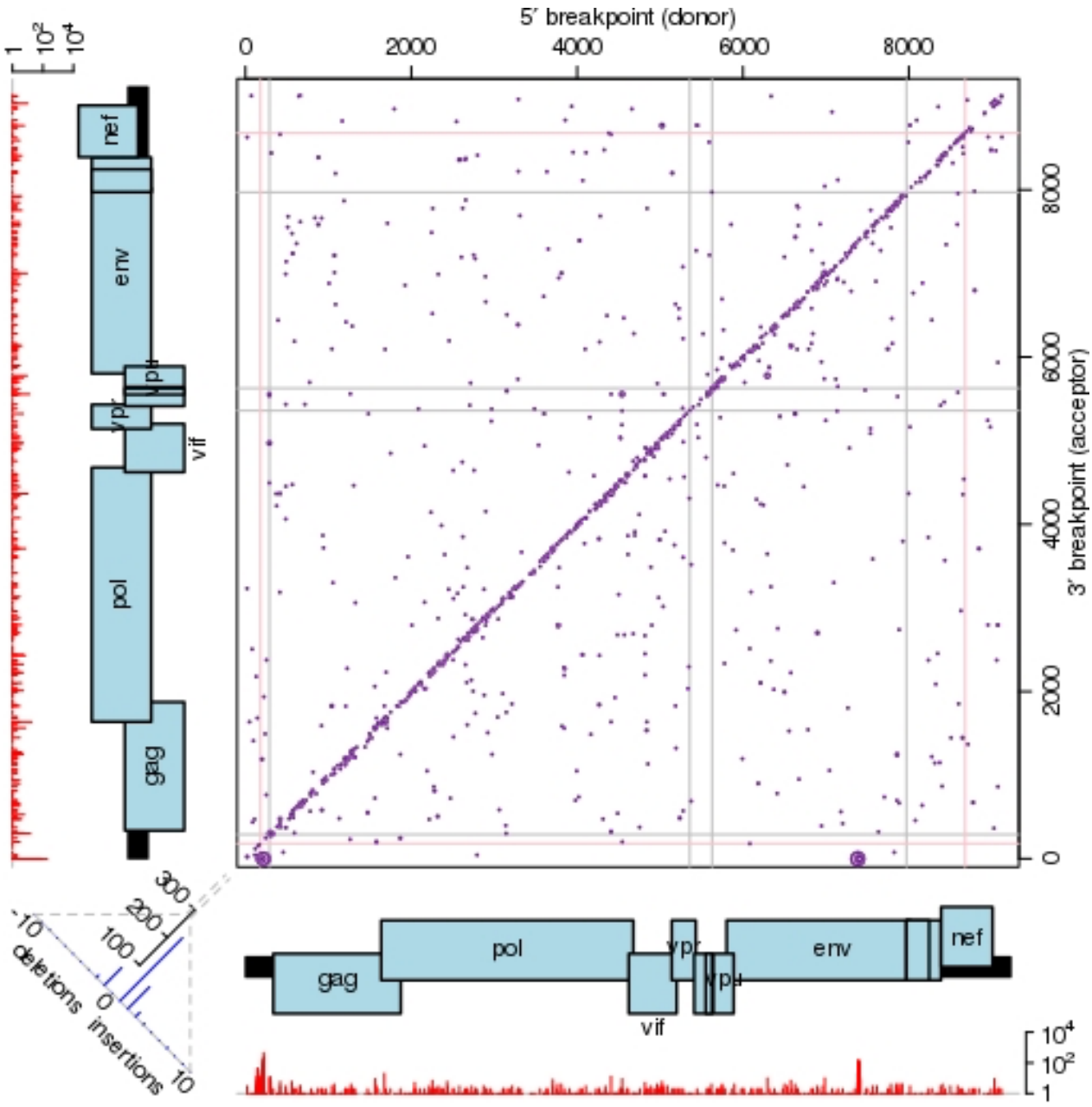


Figure 2.18 HIV chimeric reads. The horizontal axis denotes positions of the 5′ fragments of chimeric reads mapping to the HIV genome and the vertical axis denotes the positions of corresponding 3′ fragments. The area of the purple circles is proportional to the number of chimeric reads at a given position. The x-axis histogram and y-axis histogram denote the number of reads mapping to a given 5′ breakpoint and 3′ breakpoint, respectively. The purple circle in the lower-left corner corresponds to reads mapping to the splice site donor at position 290 that is shared across all fully and partially spliced HIV transcripts. The purple circle in the lower right hand corner corresponds to the unusual RNASeq peak at position 7755, as underlined by the purple dot in
Figure 2.6. Dots on the diagonal axis represent small insertions or deletions, which may occur due to polymerase slippage during reverse transcription, host RNA pol II transcription or artefacts arising during amplicon preparation and deep sequencing.

their significance it would be necessary to repeat the experiments with biological replicates, as well as potentially with different combinations of HIV-1 subtypes and cell lines (work in progress). A noted limitation of profiling is the potential for other factors, such as RBPs, to produce protected fragments that co-sediment with RPF-containing ribosomes, thus giving a misleading read-out of ribosome density. Env and Vpu are translated via ribosomes that associate with the rough endoplasmic reticulum, so one cannot rule out the possibility that the pause site seen in the *env* ORF is caused due to another co-translational mechanism. Additionally, it is possible that the localization of the ribosomes on the rough endoplasmic reticulum imposes an added spatial constraint that may affect the size distribution of RPFs on the *env* ORFs, which could be possibly lost during the gel excision stage and result in our under-estimating the size of the pause.

The study provides initial evidence for a putative uORF shared across all spliceforms in a subset of HIV-1 strains that may be involved in regulating HIV replication. Our work also constitutes a new methodological assessment of HIV frameshift efficiency in live cells and provides an estimate that is two to four times larger in magnitude than that measured using dual-luciferase constructs. Our study was limited in its ability to estimate translational efficiency due to a low level of spliced transcripts. It is likely that the majority of our viral reads were produced from newly synthesized genomic RNA. Ideally, the experiments would be repeated at several different time points post-inoculation, though such studies analyses can be challenging to perform in the laboratory. Repeating the experiments with different strains of HIV-1 would also allow one to assess whether the low level of splicing detected in our libraries was simply due to some kind of defect in the strain used in this study. Retroviruses are considered to be relatively benign at the cellular level in that even during chronic infection, only a small portion of cellular metabolism is usurped for viral reproduction and only a few percent of the cellular or organism level mRNA and protein pools is of viral origin (Goff, 2013). Our pseudo-differential expression analyses corroborated this account, as no genes were detected to have large changes in read counts between the uninfected and infected replicates, although this analysis was limited by a lack of estimates for the sensitivity of detection.

Our results demonstrate the feasibility of utilizing ribosome profiling with retroviruses, and highlights several specific caveats for future experimental work, such as the need for high-sequencing depth to enable sufficient RPF counts for analysis of retrovirus translation. We also suggest future ribosome profiling studies of retroviruses utilize longer RNASeq sequencing reads as well as potentially paired-end reads in order to maximize the chances of obtaining high

read coverage of splice boundaries that would allow for more precise detection of various isoforms (Brar et al., 2015).

In conclusion, this study provides a snapshot of the HIV translatome and the practicalities of performing ribosome profiling on retroviruses, which we hope will aid in the design of future profiling studies as well as inform studies of the retrovirus replication cycle.

Chapter Three

Computational Analyses of Coronavirus Ribosome Profiling Data

3 CHAPTER THREE: COMPUTATIONAL ANALYSES OF CORONAVIRUS RIBOSOME PROFILING DATA

3.1 Introduction

The *Coronaviridae* are a family of viruses in the order Nidovirales, consisting of the *Coronavirinae* and *Torovirinae* sub-families (Woo et al., 2009). They are of both clinical and veterinary interest due to the numerous examples of coronavirus-mediated pathologies (Table 3.1), including the heavily publicised cases of severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) (Ksiazek et al., 2003; Chan et al., 2015). For most human coronaviruses, there are currently no anti-viral therapies available, further underscoring the importance of characterizing the molecular biology of this group of pathogens (Stockman et al., 2006; Momattin et al., 2013). Previous studies have noted the complex inter-play between coronaviruses and their hosts, in which they utilise numerous proteins to modulate the immune response (Raaben et al., 2009; Fung et al., 2016).

To further elucidate the specific contribution of translation modulation to anti-virus responses, assess the relative translational efficiencies of host and virus transcripts, analyse differential uORF usage in certain genes and to investigate whether there are any virus-induced changes in gene expression which have not been previously detected, we have surveyed the cellular translatome and transcriptome throughout a time course of coronavirus infections. We utilized the murine coronavirus mouse hepatitis virus (MHV), arguably one of the most heavily studied members of the *Coronaviridae* family (Masters and Perlman, 2013). This chapter will first provide an introduction to coronaviruses and MHV, then describe a series of bioinformatics analyses to interrogate the effects of MHV infection on murine gene expression. Subsequently a series of custom visualization methods used to analyse RiboSeq and RNASeq data will be described.

3.1.1 General characteristics of coronaviruses

The *Coronavirinae* are a sub-family of enveloped, positive-stranded RNA viruses which typically exhibit groups of ~20 nm petal-shaped projections on their virions (hence the prefix ‘corona-’ to signify the resemblance with a crown). Coronaviruses have genomes which range from 26 to 32 kb in length, the largest amongst RNA viruses, and include a 5′ cap and a poly(A) tail (Sawicki et al., 2007; Smith and Denison, 2013; Hofer, 2013). Unusually for a positive-strand RNA virus, they also have helically symmetric, ~120-140 nm wide nucleocapsids and a

Coronavirus example	Tissue tropism	*Coronavirinae* genus
Human coronavirus 229E (HCoV-229E)	Pulmonary epithelium	*Alphacoronavirus*
Porcine epidemic diarrhea virus (PEDV)	GI tract epithelium	*Alphacoronavirus*
Severe acute respiratory syndrome coronavirus (SARS-CoV)	Respiratory, GI tract	*Betacoronavirus*
Mouse hepatitis virus (MHV)	CNS, liver, lungs, GI tract	*Betacoronavirus*
Middle East respiratory syndrome coronavirus (MERS-CoV)	Bronchial epithelium	*Betacoronavirus*
Bovine coronavirus (BCoV)	GI tract, respiratory	*Betacoronavirus*
Porcine hemagglutinating encephalomyelitis virus (PHEV)	Pulmonary epithelium, CNS	*Betacoronavirus*
Infectious bronchitis virus (IBV)	Respiratory, kidney	*Gammacoronavirus*
Turkey coronavirus (TuCoV)	GI tract	*Gammacoronavirus*

Table 3.1 Clinically and veterinary important coronaviruses. Coronaviruses have been shown to cause pathologies in multiple clinical and husbandry contexts (Cavanagh, 2007; Coleman and Frieman, 2014; Lau and Chan, 2015).

relatively high fidelity polymerase (Smith, 2013).

3.1.2 MHV genome and proteins

The preferred tropism of MHV is the murine brain, liver, GI epithelial and lung epithelial cells, which it infects via mCEACAM1 receptors (Hemmila et al., 2004). The MHV genome is roughly 31.3 kb in length (Figure 3.1), depending on the specific strain. The MHV-A59 genome encodes a total of 24 proteins, which are synthesized from 11 separate ORFs (Sawicki et al., 2007; Masters and Perlman, 2013). The non-structural proteins (nsp) of the *replicase* gene are translated from two ORFs located at the 5′ end of the full-length coronavirus RNA genome as parts of a polyprotein that later undergoes autoproteolytic cleavage (Brierley et al., 1987; Baker et al., 1993; Gadlage and Denison, 2010; Fung and Liu, 2014).

The *rep1a* ORF encodes the 496-kDa pp1a polyprotein, which is cleaved to form the nsp1 to nsp11 proteins. The *rep1b* ORF encodes the nsp12 to nsp16 proteins, and is expressed as part of the 802-kDa pp1ab fusion protein which is synthesized when ribosomes translating the *rep1a* ORF frameshift into the −1 reading frame (Brierley et al., 1987; Masters, 2006).
The remaining proteins are translated from downstream ORFs that are expressed from shorter sub-genomic RNAs (sgRNAs). The sgRNA are not produced through alternative splicing of the full length transcript; instead they are synthesised through a process of discontinuous negative sense synthesis where the RNA-dependent RNA polymerase ‘jumps’ between specific “transcription regulatory sequences” (TRSs) (Sawicki et al., 2007). These TRSs contain a distinctive, ~10 nucleotide AU-rich motif and are located at the 3′ end of a shared leader sequence derived from the 5′ of the genome and the 5′ end of each sgRNA body sequence. Joining of the leader and body segments of sgRNA occurs through a copy-choice like mechanism involving the TRS elements (Zúñiga et al., 2004; Yang and Leibowitz, 2015). Viral and/or cellular factors that bind to the UTRs in the genomic RNA may circularize the genome and promote template switching by the replicase complex by facilitating base pairing between the template strand TRS leader and nascent strand TRS body sequences (Sola et al., 2005; Pasternak et al., 2006). The nsp3, nsp4 and nsp6 proteins contain hydrophobic transmembrane domains that anchor the replicase-transcriptase complex (RTC) to various cellular membranes and may be involved in the mechanism of jumping (Oostra et al., 2007)

The *S* ORF encodes a 128-kDa transmembrane protein, which is processed in the endoplasmic reticulum and cleaved by a cellular protease into S1 and S2 polypeptides

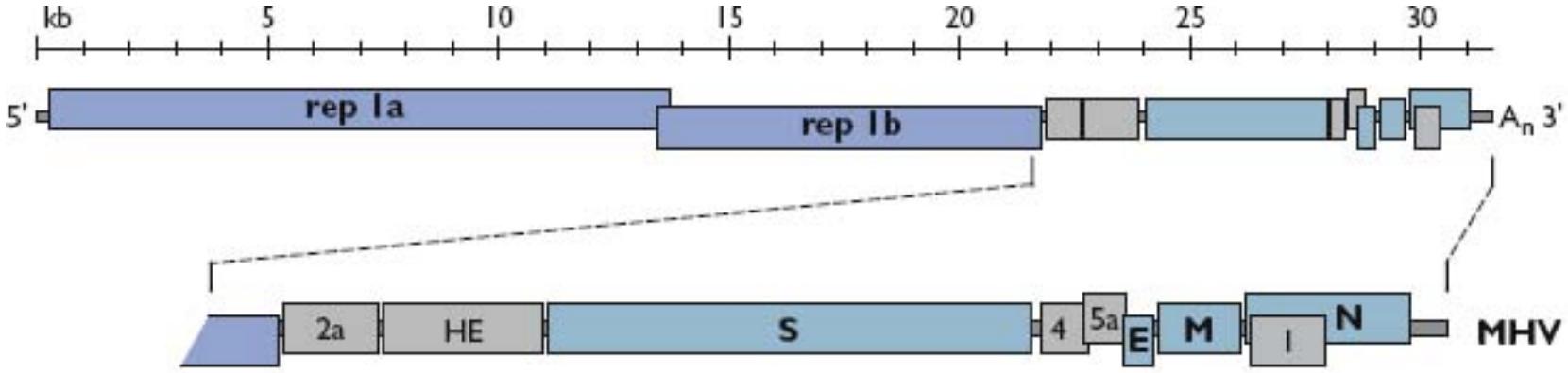


Figure 3.1 MHV genome. Illustration of a template of the generic coronavirus genome (top) and the MHV genome in particular with various MHV ORFs labelled (bottom).

(Song et al., 2004; Beniac et al., 2006). S proteins form trimer 'spikes' on the surface of the virion and facilitate fusion of the virion with the infected cell. The S1 polypeptide constitutes the receptor-binding portion of the 'spikes' protruding on the virion surface, while the S2 molecules constitute the stalk of the aforementioned spikes (Huang et al., 2006; Qiu et al., 2006; Yamada et al., 2009). The *M* ORF encodes a 25-kDa structural protein that is expressed co-translationally on the ER membrane and glycosylated post-translationally. M proteins form the coronavirus matrix and give the virion its shape. The *E* ORF is translated to produce an 8-kDa integral membrane protein that is located on the virion envelope and may function in organizing assembly and release of the MHV virion (Masters, 2006; Masters and Perlman, 2013; Fung and Liu, 2014). The *N* ORF directs the synthesis of a 43 kDa phosphoprotein that binds viral RNA and M proteins to facilitate nucleocapsid formation. It also inhibits type I interferon signalling and RNase L activity. The *I* ORF is embedded as an alternative reading frame within the *N* ORF and is translated at ~10% the level of *N*, yielding a hydrophobic polypeptide that is thought to confer a minor growth advantage (Fischer et al., 1997; Irigoyen et al., 2016). Hemagglutinin-esterase (HE) is a 48-kDA structural protein that forms sets of small projections on the surface of the virion and may act in conjunction with the S protein to facilitate attachment to cells and translocation through extracellular spaces. The HE protein has been shown to slow MHV reproduction in tissue culture (Lissenberg et al., 2005) and is not expressed in some laboratory strains. The 5a protein functions to inhibit type I interferon signalling (Koetzner et al., 2010).

In the laboratory-adapted strain of MHV-A59 used in our experiments, the ORFs encoding the 4 and HE proteins are defective.

3.1.3 MHV replication cycle

The S protein facilitates binding of an MHV virion to a target cell receptor and is the major determinant of MHV tropism (Kuo et al., 2000). It has been shown to interact primarily with murine CEACAM1 (an immunoglobulin), whose various isoforms are expressed in the mouse brain, liver, lung epithelium and GI tract epithelium. Cells that are engineered to express CEACAM1 can also be infected by MHV, and knockout mice lacking the receptor are immune to MHV infection (Hemmila et al., 2004; Hirai et al., 2010).

After the S1 peptides located on the tips of the virion 'spikes' bind, a conformational change occurs in the adjacent S2 subunits, exposing a fusion peptide that can insert into the cellular membrane and form a fusion complex (Huang et al., 2006; Qiu et al., 2006; Yamada et al., 2009). After the virus and cell membranes fuse, the virus nucleocapsid is inserted into the

cytoplasm (Figure 3.2). Some variants of MHV and other coronaviruses have been noted to use receptor-mediated endocytosis and acidic-pH triggered endosome fusion instead of plasma membrane fusion for entry into the cytoplasm (Masters and Perlman, 2013).

Once the nucleocapsid is released into the cytoplasm, the *replicase* gene on the full-length viral RNA is translated to produce several nsps that form the replicase-transcriptase complex. The *replicase* gene is expressed as either the pp1a or pp1ab fusion protein; the latter resulting from ribosomes translating the *rep1a* ORF shifting into the −1 reading frame. The MHV −1 PRF site consists of a 5′-UUUAAAC-3′ heptanucleotide slippery sequence and an RNA pseudoknot located downstream after a five nucleotide 'spacer' sequence. Various methodologies have estimated coronaviral −1 PRF efficiency to be approximately 45% (Masters and Perlman, 2013; Irigoyen et al., 2016).

The pp1a and pp1ab polyproteins undergo autoproteolytic cleavage to release the final functional viral replication proteins. The nsp3 protein contains a protease domain that separates nsp1, nsp2 and nsp3, while nsp5 has a proteolytic activity that cleaves the other 11 peptide linkers. The nsp12 peptide contains the viral RNA-dependent RNA polymerase (RDRP) that is a core component of the RTC (Fung et al., 2014).

Once the RTC has been assembled, it uses the positive-sense genomic RNA as template to produce a negative-sense RNA molecule, which is then utilized to direct synthesis of more positive-sense RNA genomic copies. The full-length negative-sense RNA is also utilized in a series of discontinuous transcription reactions, where the replicase makes a series of subgenomic RNAs that encode the various structural proteins (as discussed in section 3.1.2).

The M, S and E proteins are cotranslationally imported into the ER and are moved towards the ER-Golgi intermediate compartment (ERGIC), where they assemble alongside N protein and genomic RNA copies. ERGIC-derived vesicles move through the secretory pathway and eventually fuse with the plasma membrane, thus releasing progeny virions (Masters, 2006). In some coronaviruses, a fragment of the S protein can cause fusion of adjacent cells to form a syncytium, which provides an alternative mechanism for infecting new cells (de Haan et al., 2004; Masters and Perlman, 2013).

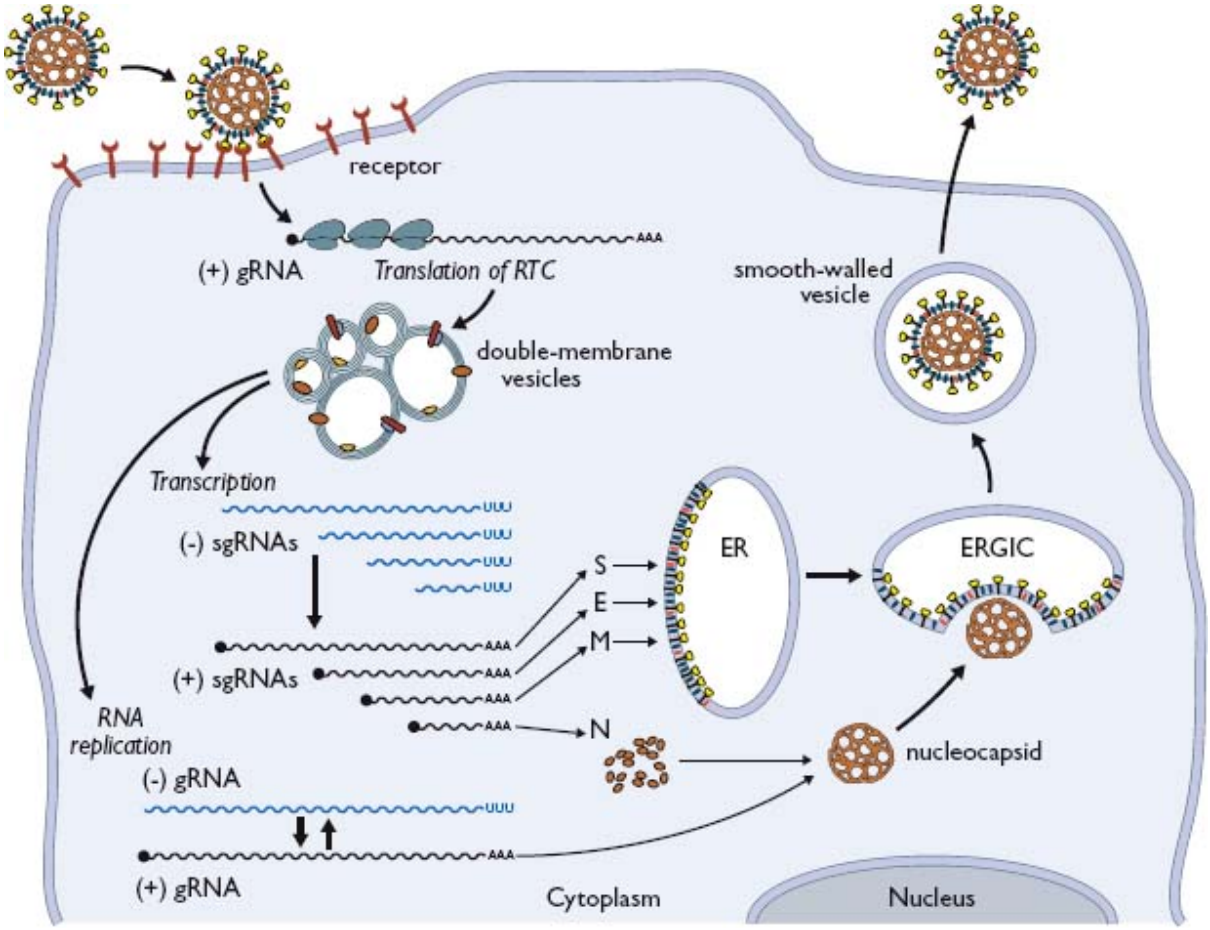


Figure 3.2 Coronavirus replication cycle. Coronaviruses replicate their single-stranded RNA genomes in the cytoplasm and make extensive use of ER/Golgi compartments for protein synthesis, modification and organization. Figure adapted from Masters and Perlman, (2013).

3.1.4 Effects of MHV infection on cellular gene expression

MHV infection results in a complex inter-play between various viral factors and cellular gene expression, as the virus repurposes the host's translational apparatus for its reproductive needs and attempts to evade the ensuing immune responses (Table 3.2).

Except for the N protein, all coronavirus structural proteins are transmembrane proteins synthesized in the ER. Previous studies have demonstrated that when protein synthesis surpasses the folding capacity, unfolded proteins accumulate in the ER and this leads to ER stress. ER stress can also be activated by excessive lipids or pro-inflammatory cytokines (Kharroubi et al., 2004; Pineau et al., 2009).

PKR-like ER protein kinase (PERK), activating transcriptional factor-6 (Atf6), and inositol-requiring protein-1 (IRE1) are transmembrane proteins that act as ER stress sensors (Figure 3.3). When unfolded proteins accumulate in the ER, the cytosolic domain of each of these sensors can trigger signalling that leads to the unfolded protein response (UPR) (Ron and Walter, 2007). As part of the UPR, cells respond by limiting protein synthesis and increasing expression of ER chaperones. The decrease in translation is mediated through phosphorylation of eIF2α, which prevents recycling of eIF2-GDP/GTP and inhibits ribosome processivity (Ron and Walter, 2007). Prolonged activation of the UPR can induce apoptosis (Tabas and Ron, 2011). Previous reports have demonstrated that coronavirus infection leads to activation of the UPR. For example, Versteeg et al. (2007) found that MHV infection upregulated homocysteine-inducible, ER stress-inducible, ubiquitin-like domain member 1 (Herpud1), an ER stress marker. Further, SARS-CoV S2 has been shown to up-regulate UPR effector-chaperones GRP94 and GRP78/BiP (Jiang et al., 2005; Yeung et al., 2008).

How each of the three UPR sensor-pathways may be influenced by coronavirus infection has been investigated in some detail. Upon induction, the IRE1 RNase domain excises a 26-nucleotide intron from the X-box binding protein 1 (Xbp1) transcript, producing a frameshifted, truncated isoform which encodes a factor that increases expression of various chaperones (Yoshida et al., 2001; Calfon et al., 2002; Fung and Liu, 2014). IRE1 also induces non-specific degradation of ER-associated transcripts through IRE1-dependent RNA decay (RIDD) (Yoshida et al., 2001; Hollien et al., 2009). Previous studies noted that MHV infection or the presence of MHV S protein induces Xbp1 splicing, but were unable to detect the spliced Xbp1 protein product in either nuclear or whole cell lysates (Versteeg et al., 2007; Bechill et al., 2008).

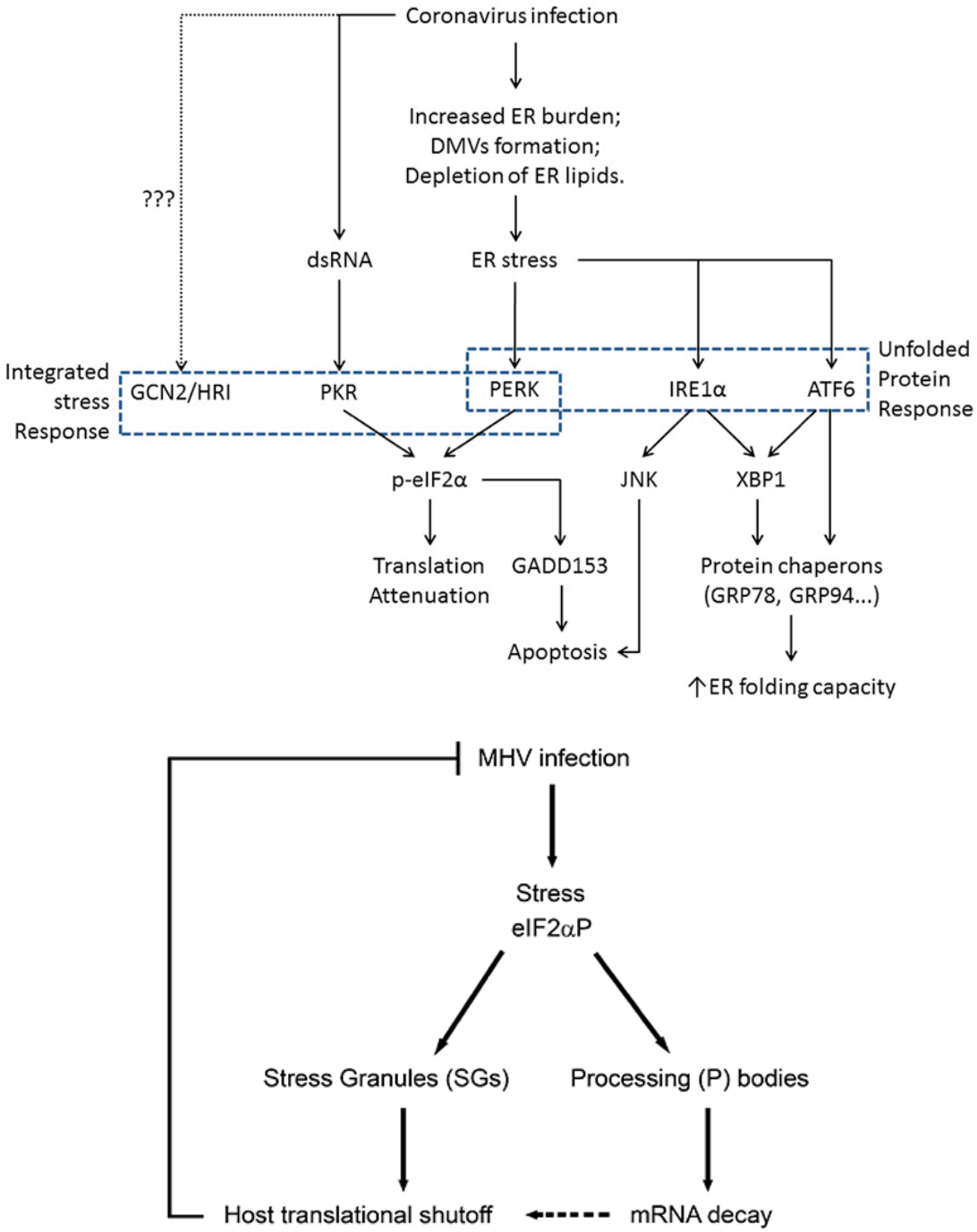


Figure 3.3 Effects of MHV infection on cellular gene expression. Top panel - MHV infection induces the Integrated Stress Response and Unfolded Protein Response. There is evidence to suggest that the UPR activation occurs though several pathways. Bottom panel - a previous study noted that MHV infection can induce stress granule and processing body formation, leading to reductions in transcript abundance for many genes. Figures adapted from Raaben et al., (2007) and Fung & Liu, (2014).

MHV Protein	Effect
nsp1	Inhibits activation of interferon signalling; suppresses host mRNA translation and/or promotes degradation of host mRNA
nsp3	Blocks interferon-mediated pathways, increases cellular pro-inflammatory cytokine expression
nsp16	Inhibits Melanoma Differentiation-Associated protein 5 (MDA5) and Interferon Induced proteins with Tetratricopeptide repeats (IFIT)
5a	Blocks interferon-mediated pathways
N	Helps block interferon-pathway activation; inhibits OAS/RNase L

Table 3.2 MHV proteins and known immune system effects. Multiple studies have examined the immune-specific effects of the MHV proteome and shown that it inhibits anti-viral response through several mechanisms (Masters and Perlman, 2013).

Fung and Liu (2014) propose that eIF2α phosphorylation and host translational repression (which are mediated through other sub-pathways) prevent Xbp1-s translation, though this has not been verified experimentally.

The Atf6 pathway increases expression of ER chaperones (GRP78, GRP94) and UPR transcription factors (GADD153, XBP1) (Szegezdi et al., 2006). GADD153 (growth arrest and DNA damage-inducible protein 153; also known as CHOP) can alter the balance between anti-apoptotic Bcl-2 and pro-apoptotic Bim as well as activate ER oxidoreductin-1α (ERO1α), leading to cell death (Marciniak et al., 2004; Puthalakath et al., 2007; Fung and Liu, 2014). Bechill et al., (2008) detected the activated Atf6 peptide 7 hours post MHV infection, but levels of cleaved and uncleaved Atf6 decreased afterwards, and there was no increase in transcript abundance of Atf6 activated genes. The authors propose that similarly to the scenario with Xbp1-s, eIF2α phosphorylation prevents translation of Atf6 and blocks activation of Atf6 target genes.

The PERK branch of the UPR pathway is responsible for mediating phosphorylation of eIF2α, and is the first branch activated upon elevation of ER stress (Szegezdi et al., 2006; Ron and Walter, 2007; Fung and Liu, 2014). Despite eIF2α phosphorylation-induced translational shutoff, certain mRNAs that contain uORFs in their 5′ UTR are preferentially translated under these conditions (Fung and Liu, 2014), and some viruses, despite contributing to ER overload, have evolved counter-measures to prevent PERK/eIF2α from arresting cellular translation (Pavio et al., 2003; Mulvey et al., 2007). There have been conflicting reports on the role of the PERK pathway in the context of coronavirus infection. Ye et al. (2007) did not detect phosphorylated eIF2α in cells infected with MHV-A59, while two other studies did (Versteeg et al., 2007; Bechill et al., 2008). The latter study also detected sustained translational suppression and upregulation of Atf4, but did not observe induction of Ddit3. Fung and Liu (2014) propose that the seemingly conflicting results in these reports may be due to the use of different cell culture systems and/or virus strains.

In addition to PERK, three other kinases are known to phosphorylate eIF2α in response to various stressers – protein kinase, RNA-activated (PKR), heme-regulated inhibitor kinase (HRI) and general control non-derepressible-2 (GCN2) (Ron and Walter, 2007). These kinases represent the sensors of the integrated stress response (ISR) (Figure 3.3). Relevant to the present discussion is the ability of PKR to be activated by interferon signalling and the presence of dsRNA. PKR is thus a component of the viral immune response, and previous studies have shown that some viruses have evolved mechanisms to subvert PKR signalling (He, 2006). Whether the ISR or the UPR is pre-dominantly responsible for phosphorylation of eIF2α in

response to MHV infection is uncertain. Given that Zorzitto and colleagues (2006) observed minimal activation of PKR and the interferon stimulated gene OAS in MHV-1 infected lung fibroblast cells, while several other studies detected both ER overloading and PERK phosphorylation in response to MHV infection, it has been suggested that eIF2α-mediated translational attenuation is primarily a result of the PERK/UPR pathways and not a direct result of the coronavirus activating PKR and the ISR (Fung and Liu, 2014).

In addition to utilizing the ER for protein synthesis and post-translational modifications, it has been shown that the replication of many plus-stranded RNA viruses induces modification of cellular membranes, which may further perturb cellular homeostasis (Miller and Krijnse-Locker, 2008). Reggiori et al. (2010) noted that MHV infection can cause an accumulation of ER-derived vesicles and associated proteins, EDEM1 and osteosarcoma amplified 9 (OS-9), in proximity to MHV replication complexes.

In summary, coronavirus infection results in large-scale morphological rearrangements of the ER, overloading of ER protein synthesis and folding pathways, and depletion of ER lipid components. MHV transcripts appear to be resistant to the effects of eIF2α phosphorylation, though the mechanisms responsible for this has not been established. We aimed to use RiboSeq and RNASeq to survey the roles of the various UPR/ISR branches in response to MHV infection as well as to investigate whether MHV transcripts have any specific features that allowed them to be preferentially translated during shutoff.

3.2 Ribosome Profiling of MHV infected cells

To study the kinetics of virus RNA and protein synthesis, Irigoyen and colleagues (2016) performed ribosome profiling on MHV A59-infected murine 17 clone 1 (17Cl-1) cells harvested at 1, 2.5. 5 and 8 hours post-infection (h p.i.), as well as mock-infected samples at 1 and 8 hours. For each time point and condition, two biological replicates were prepared. 5 minutes prior to harvesting, cells were treated with harringtonine and/or cycloheximide as discussed in Chapter 1. Both RiboSeq and RNASeq libraries were prepared for each time point. RNASeq samples were treated with RiboZero and RiboSeq samples were treated with duplex-specific nuclease (DSN) to deplete rRNA. Amplicon libraries were sequenced using an Illumina HiSeq 2000 (repeat 1 at the Wellcome Trust Centre for Human Genetics - Oxford Genomic Centre; repeat 2 at the Beijing Genomics Institute).

Reads were trimmed for adaptor sequences, filtered, and mapped to *Mus musculus* rRNA (allowing for two mismatches). Remaining reads were aligned directly to the mouse genome

(UCSC, assembly mm10) using TopHat (parameters: --no-novel-juncs --bowtie1 --prefilter-multihits --max-multihits 500, with --transcriptome-index defined using the genes.gtf file from the UCSC mm10 annotation available from the tophat website) (Trapnell et al., 2009).

Irigoyen et al. (2016) checked read length distributions, histograms of the positions of host reads relative to start and stop codons, and read framing distributions to assess the quality of data. Analyses of the kinetics of virus replication indicated that at 2.5 h p.i., there was little change in positive-sense RNA, while negative-sense MHV RNA rose to about 0.1% of total cellular mRNA and virus RNA. At 8 h p.i., the relative proportion of positive-sense virus RNA rose to 80-90%. Viral transcripts did not have a higher translational efficiency than host mRNAs, which (together with the previous analysis of mRNA abundances) suggested that MHV is able to usurp the host translational apparatus not through preferential translation but rather through high levels of transcription.

Comparison of the RNASeq read coverage densities and leader/body chimeric reads indicated that at 5 h p.i., the relative abundances of various MHV sgRNAs was mRNA7 > mRNA6 > mRNA1/mRNA5/mRNA3 > mRNA4/mRNA2. Irigoyen et al. (2016) also utilised their RiboSeq data to confirm that the previously described internal (I) ORF that is embedded within the N gene is translated at about 10% the level of N (Fischer et al., 1997). The production of this protein was confirmed by Western blotting. RPF data supported translation of an AUU-initiated 15-codon ORF upstream of ORF4a, a CUG-initiated 13-codon ORF upstream of ORF5, a UUG-initiated single-codon site in the leader sequence, and a AUG-initiated uORF on the gRNA.

Close inspection of the RiboSeq data showed that (contrary to expectations) ribosomes do not pause at the MHV −1 PRF site between ORF1a and ORF1b. A large pause site was observed in the sequence encoding nsp3, and was independently validated in an in vitro translation system. This pause occurred in the linker region between the ADRP and DPUP domains of nsp3, and may function to allow proper protein folding (Chen et al., 2015b; Neuman et al., 2008). This chapter describes a continuation of this project, with a focus on analysing the ribosome profiling and RNASeq data to study the effects of MHV infection on murine gene expression.

3.2.1 Differential expression analyses

Reads from Irigoyen et al. (2016) were tabulated using htseq-count (parameters: -t CDS -m intersection-strict -i gene_id -s yes) with the NCBI RefSeq mRNAs as the gene feature annotation. For differential expression analyses, the htseq-count parameter "-type CDS" was used to count only those reads that were mapped to the annotated, coding portions of transcripts (Anders et al., 2015). Thus the differential expression analyses exclude reads mapping to uORFs or non-annotated coding sequences (unless these sequences overlapped with the main annotated ORF). Digital read count data from RiboSeq and RNASeq data were analysed with a variety of methods. Although, nominally, the various publicly available differential expression software have the same purpose, they each have various statistical and programming idiosyncrasies (Rapaport et al., 2013; Soneson and Delorenzi, 2013; Zhang et al., 2016). The analysis of ribosome profiling data introduces additional challenges due to the paired nature of dual-assay designs as well as the fact that the mapping space for RPFs is a distinct subset of the mapping space for RNASeq reads in a given biological replicate. Previously published profiling studies have analysed differential translation data using edgeR (Rooijers et al., 2013; Jensen et al., 2014) or edgeR and DESeq2 (Vasquez et al., 2014). Using pipelines optimised for RNASeq differential expression analyses directly for ribosome profiling presents some problems, as the RNASeq read counts have to be limited to the same regions as that for the RiboSeq analyses (i.e. coding sequences) and quotients for RiboSeq/RNASeq for translational efficiency calculations have less amenable statistical distributions (Cauchy instead of negative binomial or Poisson distributions).

Differential expression analyses on RNASeq count data or RiboSeq data were performed with DESeq (Anders et al., 2012), edgeR (Robinson et al., 2010), NOISeq (Tarazona et al., 2015), and BaySeq (Hardcastle and Kelly, 2010). Read counts were normalized by library size prior to comparison and low count genes were discarded (i.e. transcripts with fewer than 10 reads mapping on average in the analyzed libraries). For BaySeq analyses, the sample size used to calculate priors was set to 200000. For each comparison, two replicates from infected cells were compared to four libraries from uninfected cells (1 and 8 h; two replicates each). Changes in translation activity were estimated using Bayseq in the paired library mode as well as Babel (Olshen et al., 2013) and Xtail (Xiao et al., 2016). A given gene was considered to be differentially expressed if the calculated false discovery rate (FDR) value was less than 0.05 and the absolute value of log2(fold change) between the averages of infected and mock replicates was greater than 1. Volcano plots and inter-replicate consistency plots were generated using standard R plotting features and the calculated FDR and log2(fold change) values from the BaySeq, Babel and Xtail analyses. During the course of our analyses, I found several errors

in the Xtail source code regarding reporting of estimated translational efficiency ratios and p-values, which I corrected and reported to the authors of the package; my proposed modifications were pulled into the public Xtail repository (*https://github.com/xryanglab/xtail/commits/master*).

The various RNASeq analyses had a large number of matching predictions with regards to differentially expressed genes, though the non-parametric NOISeq model called 81 genes as being significantly upregulated according to our thresholds that were not called by the other three packages (Figure 3.4). The Paired BaySeq, Babel and Xtail analyses had few overlapping predictions of genes that were significantly differentially expressed at the level of translational efficiency, with Xtail producing a significantly higher number of predictions than the other two packages (Figure 3.4).

3.2.2 Effects of MHV on host translation

As discussed earlier, MHV infection is known to cause ER stress, leading to induction of the UPR and phosphorylation of eIF2α. Ribosome profiling enabled us to survey the changes in host translation on a genome wide scale and to dissect the modulation of interferon-signalling, UPR and other stress-induced genes.

Consistent with previous studies on phosphorylated eIF2α-resistant translation (Jousse et al., 2001; Lu et al., 2004; Vattem and Wek, 2004; Watatani et al., 2008; Zhou et al., 2008; Chen et al., 2009; Lee et al., 2009; Hatano et al., 2013; Andreev et al., 2015), the differential expression results detected Atf4, Atf5 and Ddit3 as being significantly upregulated at the translational level at 8 h p.i. (

Figure 3.5, Table 3.3). The *Atf4* and *Atf5* transcripts contain multiple uORFs in their 5′ leaders and undergo selective delayed re-initiation. Under conditions of low eIF2 availability, there is an increased probability of the 40S ribosome scanning past the second intervening uORF and reaching the main *Atf4/Atf5* CDS. *Ddit3* expression shows the feature of alleviation of scanning ribosome obstruction, where a decrease in loading lowers the number of ribosomes initiating at the uORF and thus decreases the obstruction of scanning ribosomes which are in turn able to proceed to the main ORF (Palam et al., 2011; Andreev et al., 2015). The use of these two mechanisms may be a large determinant in the ability of these genes to undergo selective changes in translation efficiency during MHV induced eIF2α-phosphorylation and general translation suppression.

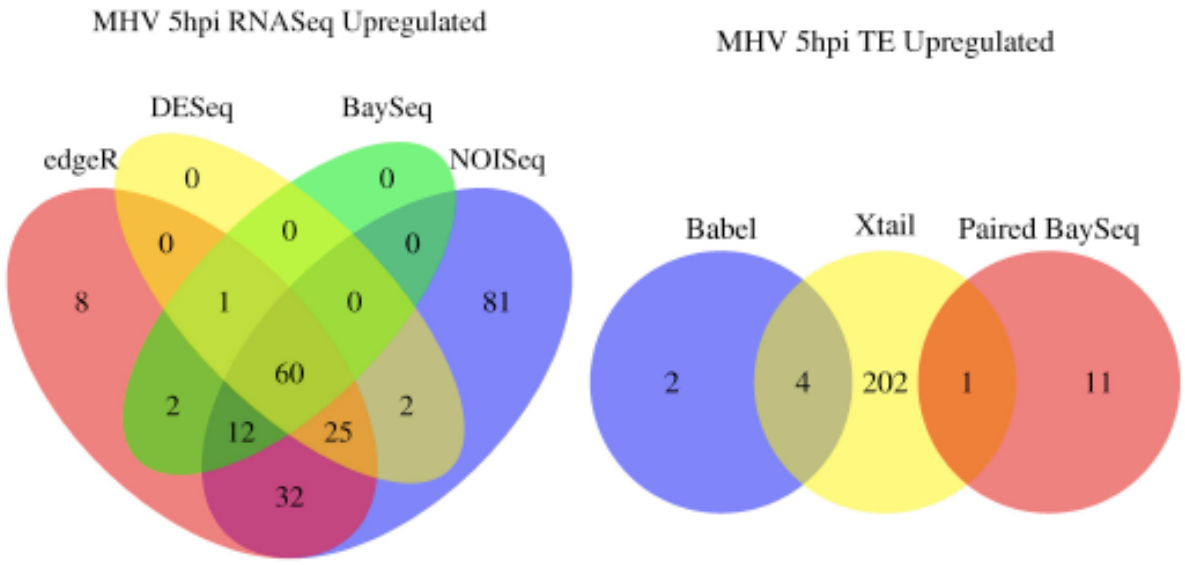


Figure 3.4 Detection of differentially expressed genes by various packages. (A) Left panel shows the number of genes identified by each package for which there was at least a two-fold increase in RNASeq read abundance after MHV infection at an FDR threshold of 0.05. (B) The right panel shows the number of reads for which the RiboSeq/RNASeq count ratio increased at least two-fold in response to MHV infection as detected by each package, with an FDR cut-off of 0.05.

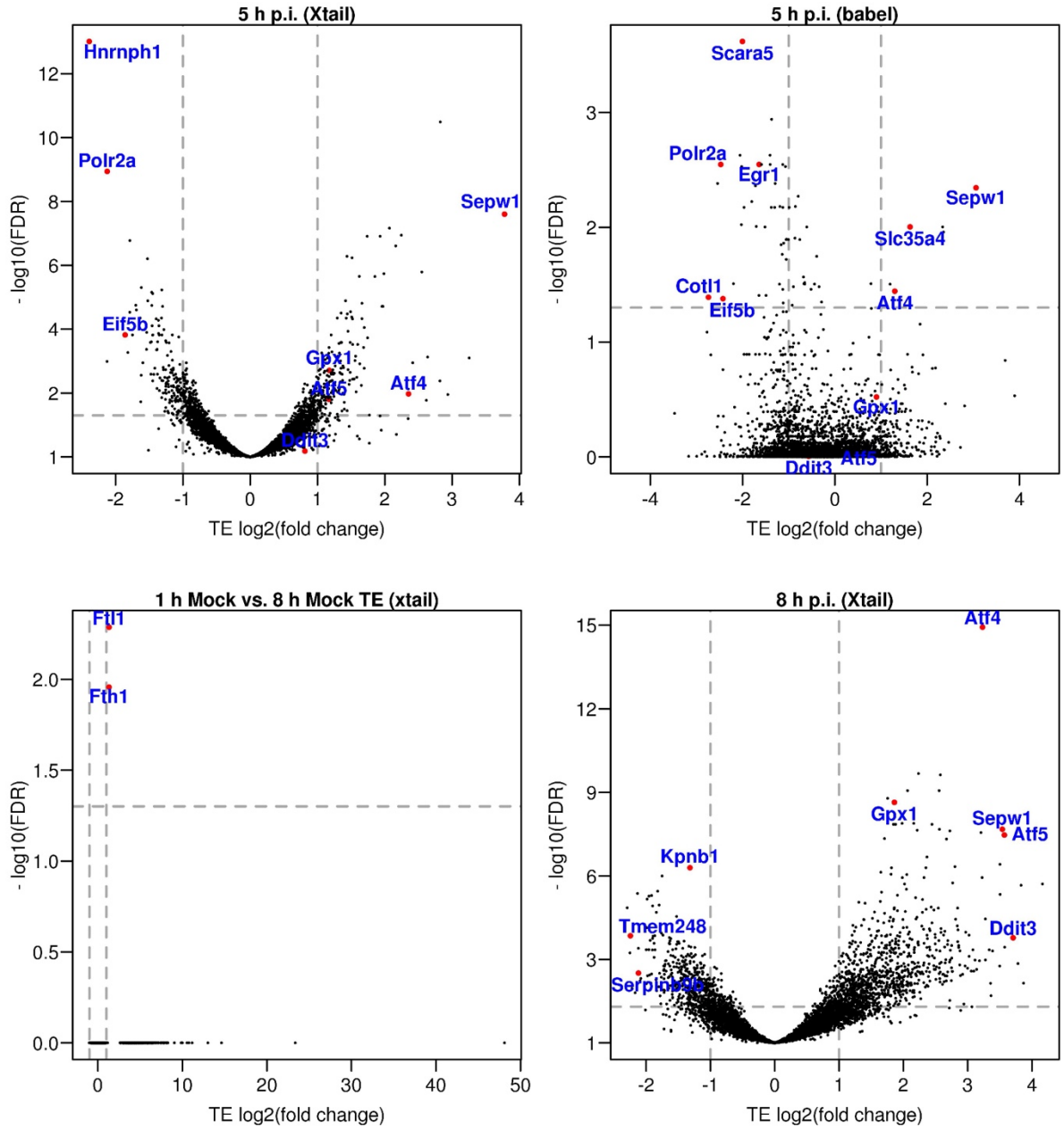


Figure 3.5 Volcano plots of differentially translated genes. Top left panel shows differentially translated genes at 5 h p.i., as detected by Xtail. Top right panel shows the same analysis but using Babel instead of Xtail. Bottom-left panel was a cell-cycle control – i.e. a comparison of the 1 hr and 8 hr mocks to see if any genes were detected as differentially translated due to changes in the cell-cycle or random variance. The two genes detected, ferritin light chain (Ftl1) and ferritin heavy chain (Fth1) are components of a protein that stores iron ions. Both these transcripts contain iron-responsive elements (IRE), RNA secondary structures that interact with iron response proteins under conditions of low iron abundance in order to inhibit translation of ferritin. Differential translation of these transcripts is suggestive of depletion of some nutrients in the media at later time points (independent of virus infection). Bottom left panel shows differentially translated genes at 8 h p.i., as detected by Xtail.

Gene	log2FC TE	FDR
Atf4	3.230825474	1.18081E-15
Rrp7a	2.555248623	8.61294E-10
Rps25	2.064275584	8.61294E-10
Rpl23a	1.754805022	1.63426E-09
Gpx1	1.860042849	2.27892E-09
Sepw1	3.538548829	2.11459E-08
Fth1	2.560478321	2.11459E-08
Atf5	3.570151208	3.40573E-08
Hmgxb4	3.501497215	3.82993E-07
Traf3	4.162940549	1.94441E-06
Eif2ak4	2.029729273	1.89941E-05
Mapk7	1.798008402	3.06048E-05
Ppp1r15a	3.274654366	3.53923E-05
Slc27a1	2.413501598	9.23387E-05
Ddit3	3.706200711	0.000168219
Slc25a37	1.787555589	0.000304994
Dicer1	2.053715081	0.000308397
Ftl1	1.948557516	0.000638044
Slc25a28	1.33934637	0.000658897
Mapk8ip3	2.728380858	0.001045213
Mapkapk5	2.688531071	0.002810501
Rps29	1.899584573	0.005997074
Golga4	2.632855095	0.006321307
Eif2ak3	1.60938558	0.006634028
Map4k4	2.383972814	0.00928179
TIA	1.015771	0.018305
Map7d1	1.71212809	0.020525764
Src	1.232319008	0.025453513
Eif2s2	1.273511337	0.030363081

Gene	log2FC TE	FDR
Sepw1	3.779308815	2.50921E-08
Rpl29	1.660856246	1.5441E-05
Gpx4	1.294139506	4.47222E-05
Eif4e2	1.021075578	0.00195662
Gpx1	1.175557763	0.00195662
Rpl41	1.331204168	0.002289284
Rps28	1.225239475	0.002298167
Slc25a37	1.26977754	0.008235743
Atf4	2.353070695	0.01063255
Atf5	1.170629408	0.015797352
Ftl1	1.771974436	0.049134409

Table 3.3 Translationally upregulated genes after MHV infection. Selection of genes with an increase in translational efficiency after MHV infection, as detected by Xtail with an FDR cut-off of 0.05. Left panel is 8 h p.i.; right panel is 5 h p.i.

The interferon-modulated developmental regulator *Irfd1* was also translated more efficiently. One isoform of this gene contains a 51-codon uORF that normally mediates mRNA decay but allows translation of the main ORF when the UPR is triggered (Gu et al., 2009; Zhao et al., 2010). In agreement with a previous analysis, we did not observe a change in *Irfd1* transcript abundance at either 5 h p.i. or 8 h p.i (Andreev et al., 2015a). The UPR has been noted to upregulate the expression of several solute carrier proteins, possibly to counteract imbalances in ion or metabolite concentrations (Dejima et al., 2009; Epple et al., 2013; Chen et al., 2016). We observed an increase in translation of a subset of solute carrier family proteins, including Slc39a13 (zinc), Slc25a37 (iron), Slc27a1 (long-fatty-chain) and a possible increase in the case of Slc35a4 (though the last one only in the Babel analysis and did not reach statistical significance).

In addition to the modulation of the UPR and interferon-signalling pathways, two components of the ISR – GCN2 and PERK, had an increase in translation efficiency.

Several mitogen activated kinases (MAP) were also translationally upregulated at 8 h p.i. MAP kinases regulate several pathways, including apoptosis, innate immunity and pro-inflammatory response (Keshet and Seger, 2010). There is a complex inter-play between these enzymes and an infecting coronavirus – it is known that MHV infection upregulates activity of certain MAPs (McGilvray et al., 1999; Xu et al., 2014; Fung et al., 2016), though there have also been reports that as the infection progresses, there is a reduction in MAP p38 activation and related drop in IL-6 secretion (Banerjee et al., 2002). Some of the MAP proteins are known to interact with various tumour necrosis factor receptor associated factors (TRAFs), which mediate apoptosis and anti-viral responses (Häcker et al., 2011). *TRAF3* had the largest change in translation efficiency, with a ~17.8 fold increase at 8 h p.i. TRAF3 is an adapter protein that has several distinct roles for different receptors in the same cell, and also as well as cell-type specific functions (Yi et al., 2014). The role of this gene in the context of MHV infection is not known.

We observed changes in the expression level of several miscellaneous genes which have previously been reported to utilize alternative initiation mechanisms or regulate translation, though the exact role of these genes in the context of MHV infection is not known. Src, a proto-oncogene tyrosine-protein kinase which was previously reported to undergo eIF2-independent translation and potentially harbour an IRES (Allam and Ali, 2010),was also found to be upregulated translationally ~2.3 fold. Interestingly, eIF2S2, the subunit of eIF2 whose catalytic activity allows the GDP-GTP exchange for successive rounds of initiation and is blocked under conditions of eIF2α phosphorylation, had a ~2.4 fold increase in normalised RPF levels. A

previous report showed that MHV-infection and eIF2α-phosphorylation in LR-7 cells results in the aggregation of RNA binding protein T-cell internal antigen-1 (TIA) and formation of stress granules (SGs; Raaben et al., 2007). TIA can cluster in a prion-like manner and the resulting SGs sequester transcripts with stalled pre-initiation complexes to regulate their translation and sort them for storage, degradation or further translation (Gilks et al., 2004; Anderson and Kedersha, 2006; Raaben et al., 2007). Our analyses identified TIA as significantly translationally up-regulated ~2 fold at 8 h p.i. Whether TIA and the aforementioned genes are upregulated due to an antiviral response or as a strategy to facilitate viral replication is not known.

3.2.3 Effects of MHV on host transcript abundance

Though eIF2α-phosphorylation decreases overall levels of translation, induction of the UPR is known to cause an increase in activity of certain transcription factors. We sought to assess whether the MHV infection caused changes in gene expression that occurred primarily at the level of transcript abundance rather than due to changes in translation efficiency. The BaySeq RNASeq differential expression analysis identified 116 transcripts with reduced abundance at 5 h p.i. and 1065 at 8 h p.i, and 60 upregulated at 5 h p.i. and 996 at 8 h p.i. *Polr2a* was found to be transcriptionally upregulated ~5.2 fold, but had a ~4.34 fold decrease in translational efficiency (Figure 3.6, Table 3.4). The ER stress marker *Herpud1* was transcriptionally upregulated ~5.1 fold, but had no significant change in translational efficiency (Figure 3.7). *Eif1a* was transcriptionally upregulated but had no significant change in translational efficiency. Changes in the transcriptional abundance of these transcripts were experimentally validated by qRT-PCR (Dr Nerea Irigoyen*, personal communication*).

Several groups have shown that Rpl19 expression is not affected by ER stress, and it has been considered a 'housekeeping gene' for normalizing gene expression studies (Hollien and Weissman, 2006; Hiramatsu et al., 2011). However, in the present study, *Rpl19* was shown to be moderately downregulated relative to the global mean after 5 h p.i. at the transcript level, though this could be due other reasons besides ER stress such as depletion of host tRNAs and altered transcription, or differences in methodologies used to quantify Rpl19 transcript abundance (Figure 3.6).

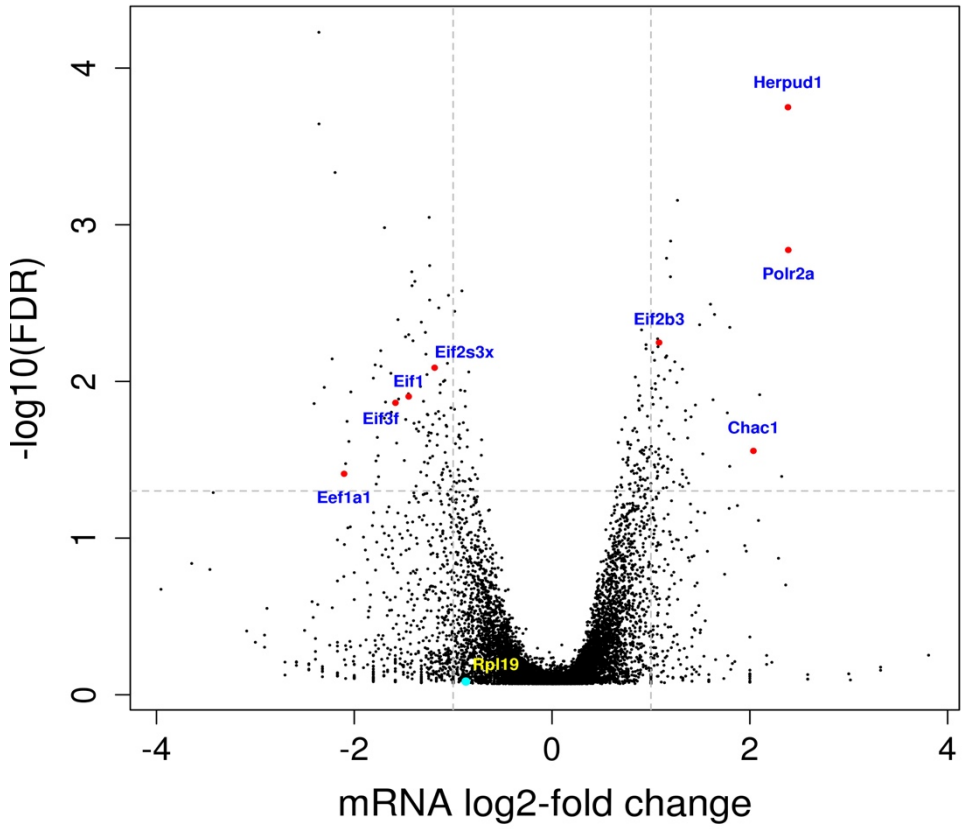


Figure 3.6 Genes with changes in transcript abundance. Genes were considered to be significantly modulated at the transcript level if there was log2 fold change greater than 1 or less than -1 in RNASeq read counts between infected and control replicates, and a FDR value less than 0.05, as estimated by BaySeq. Different components of the translational apparatus were upregulated or downregulated in this analysis. Rpl19 has been considered to be a 'housekeeping' gene and used for normalizing gene expression studies, but our data suggest that this gene is moderately downregulated (relative to the global mean) after 5 h p.i. at the transcript level.

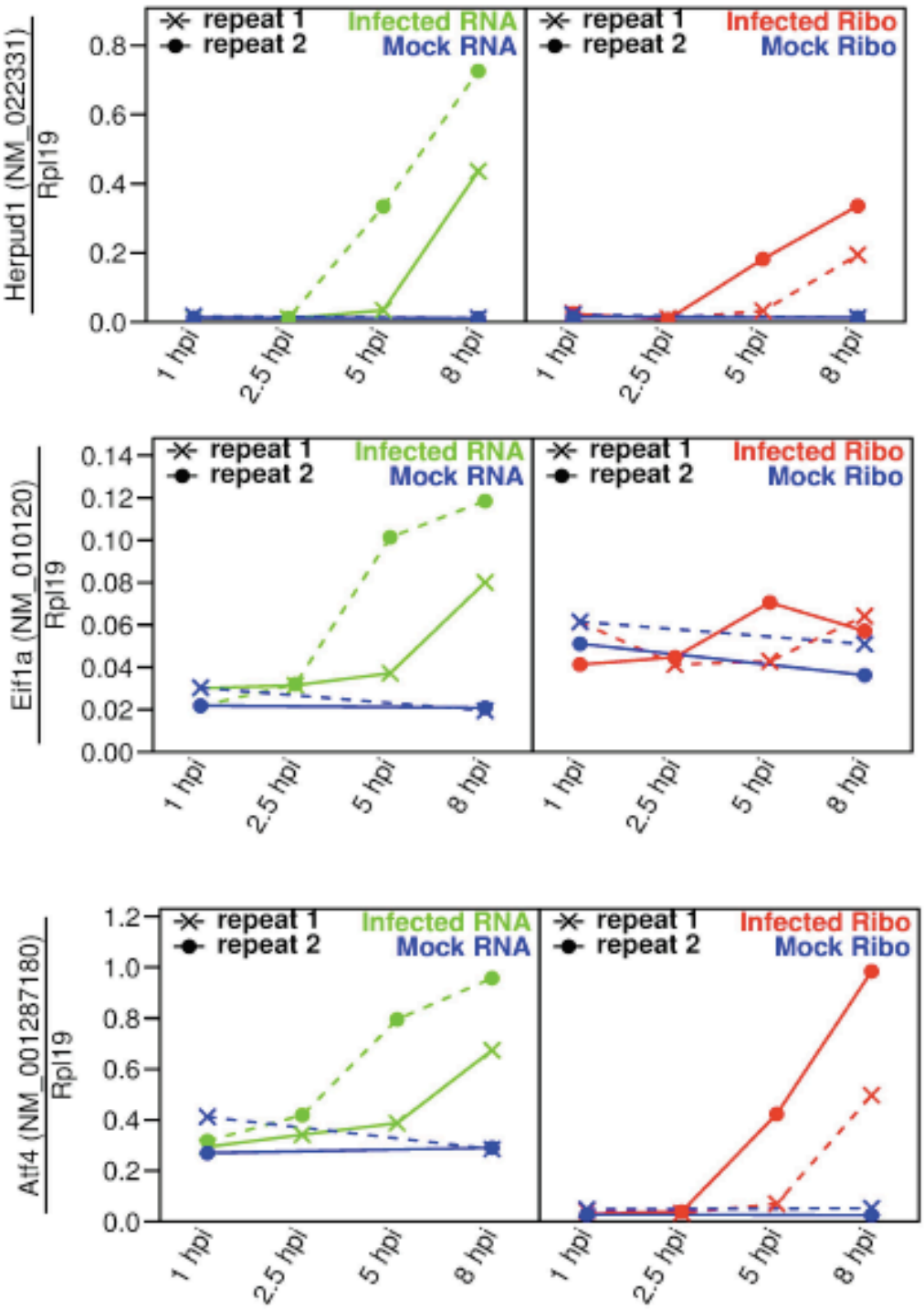


Figure 3.7 Time course of specific mRNA transcript abundances after MHV infection. RNASeq and RiboSeq counts from several differentially expressed genes, normalized by the respective values for Rpl19 (for comparison with the qRT-PCR results). Herpud1 and Atf4 are up-regulated transcriptionally and translationally as MHV infection progresses, while Eif1a is up-regulated primarily at the transcriptional level. Note that given the fluctuations in Rpl19 levels at later time points, the 5 h p.i. and 8 h p.i. time points underestimate the ratios relative to the global mean.

Gene	log2FC	FDR
Rpl29	-2.367116869	2.64001E-05
Herpud1	2.378755239	0.000106475
Polr2a	2.387461255	0.001029973
Rps6	-2.040131169	0.010004611
Eif1	-1.453638267	0.010298073
Ddit3	1.04508789	0.024986674
Vegfa	1.164494038	0.033888377

Gene	log2FC	FDR
Txnip	6.578995659	5.0413E-12
Rps6	-2.95573366	8.164E-09
Herpud1	3.691651701	1.74802E-08
Gadd45a	3.906890596	1.0792E-07
Ddit3	2.287281952	2.1342E-06
Rpl29	-2.675116938	2.4463E-06
Vegfa	2.263962482	2.59143E-06
Fth1	-3.390872078	2.73024E-06
Irf1	3.137503524	3.45379E-06
Cxcl10	5.114089343	4.4124E-06
Apol9a	3.088593923	6.52299E-06
Hsp90b1	1.987222234	1.93432E-05
Hspa5	1.638982952	0.000447455
Polr2a	3.131665668	0.001126006
Hspa9	1.436944755	0.002344942
Dusp1	2.087031993	0.002432225

Table 3.4 Genes with changes in transcript abundance after MHV infection. Selection of genes that were identified as differentially expressed in the RNASeq analysis. Left panel is 5 h p.i.; right panel is 8 h p.i.

3.2.4 Visualising RPF alignments

In addition to determining the total read counts for each host gene, we also visualised how reads mapped to specific regions of given transcripts, including uORFs. As reads were mapped against the genome rather than against the transcriptome, genomic read mapping locations had to be converted to transcript-specific coordinates before generating plots showing distributions of RNASeq and RPF reads. After aligning reads using tophat, the resulting bam files were sorted by genomic coordinates and indexed using SAMTools (Li et al., 2009b). A custom R script using Rsamtools (Delhomme et al., 2012) was developed to quickly extract reads at a given series of genomic loci from a bam file (corresponding to the exon sequences of a given gene isoform), convert the alignment positions of these reads to positions relative to the transcript sequence, calculate the phasing of the reads and generate plots showing the distributions of the reads. In some cases it is not possible to distinguish usage of alternative exons or changes in transcription start sites with changes in initiation codon utilization. The advantage of the aforementioned visualization process is that it retains all RiboSeq reads mapping to a given exon and allows manual inspection of data which may originate from different isoforms. Read positions for RPFs were offset +12 nt so that RPFs whose 5′ end aligned to the first position of a codon were mapped to the first nucleotide of the inferred P-site codon, and RPFs whose 5′ end aligned to the third position of a codon were mapped to the last nucleotide of the codon preceding the P-site codon.

We scrutinised the codon-specific translation patterns of a variety of stress-response genes and observed some interesting features. Consistent with previous studies on the regulation of Atf4 in the context of the UPR, we saw a decrease in the number of RPFs to the second *Atf4* uORF as MHV infection progressed, and a concomitant increase in the number of RPFs mapping to the main *Atf4* CDS (

Figure 3.8). *Atf5* exhibited a similar decrease in uORF RPFs and an increase in translation of the main ORF (

Figure 3.9). For *Atf6*, we noted some RPFs mapping to the transcripts at 8 h p.i. and while the relative change between mock and infected replicates seems minimal, the absolute number of RPFs mapping to the transcript in both conditions was low making it difficult to determine whether Atf6 is being activated (

Figure 3.10). In the case of *Ddit3,* we were surprised to see that an increase in the number of RPFs mapping to the main ORF was accompanied by an *increase* in the number of RPFs

from initiating ribosomes on the *Ddit3* uORF (

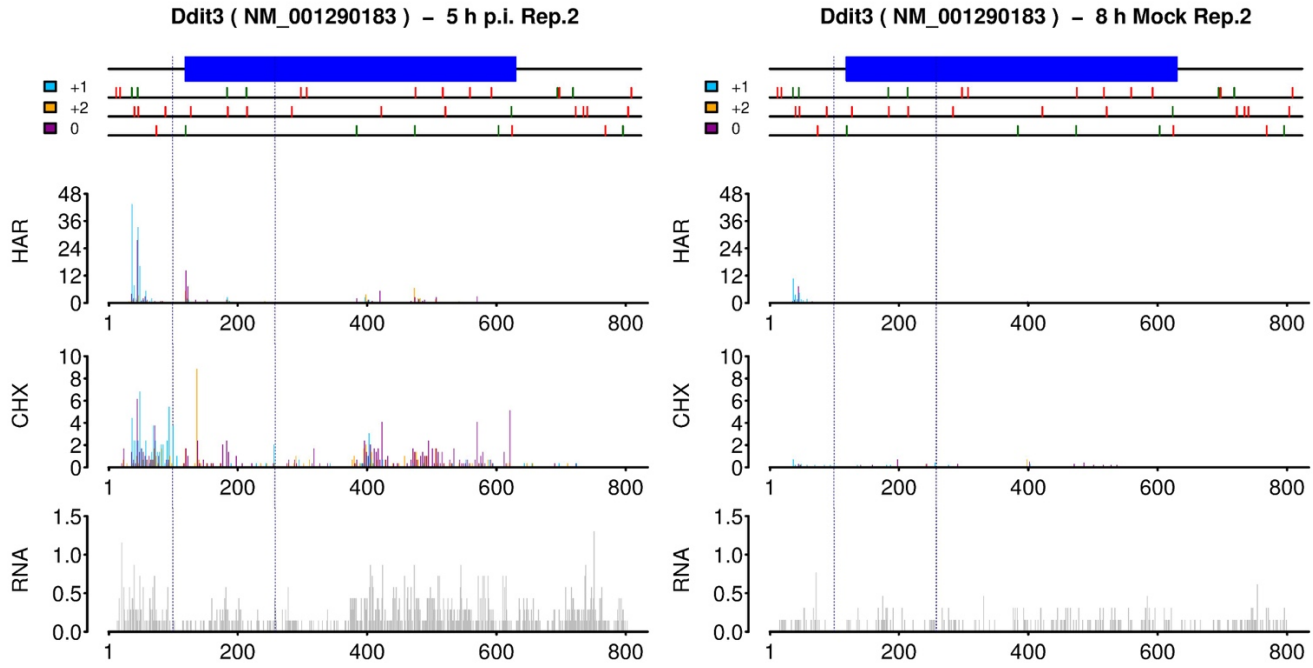


Figure 3.11), in contrast to the model described by Chen et al. (2009) and Palam et al. (2011). In one of the 8 h p.i. replicates, we also observed a peak corresponding to a downstream AUG located in the *Ddit3* ORF, suggesting the possibility

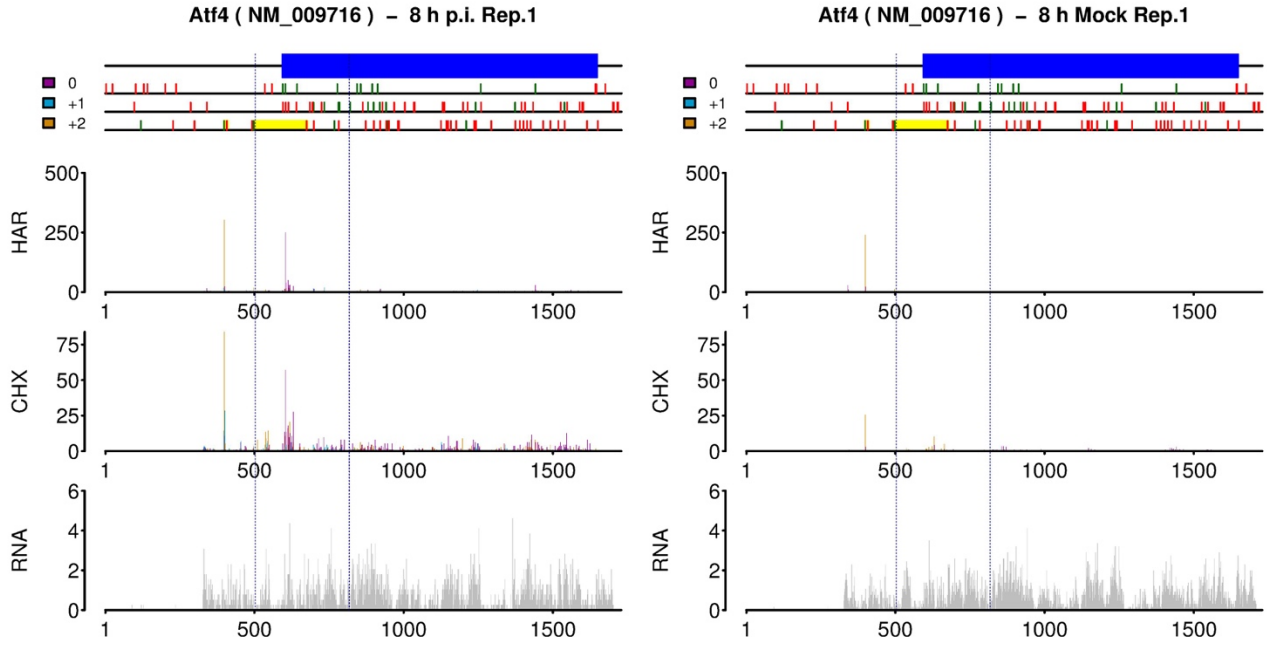


Figure 3.8 Atf4 RiboSeq plot. The blue rectangles indicate the annotated Atf4 CDS and yellow rectangles indicate Atf4 uORFs. The green tick marks correspond to AUG codons and the red tick marks correspond to UGA, UAG and UAA codons. The purple box indicates the 0-frame

relative to the annotated start codon, the cyan box indicates the +1 reading frame and the golden box indicates the +2 reading frame. Blue dashed vertical lines indicate annotated exon boundaries. The left panel indicates read distributions at 8 h p.i. while the right panel indicates read distributions in the 8 h mock control. The top pair of plots are from harringtonine treated libraries, the middle pair of plots are from cycloheximide treated libraries and the bottom pair of graphs are from RNASeq libraries.

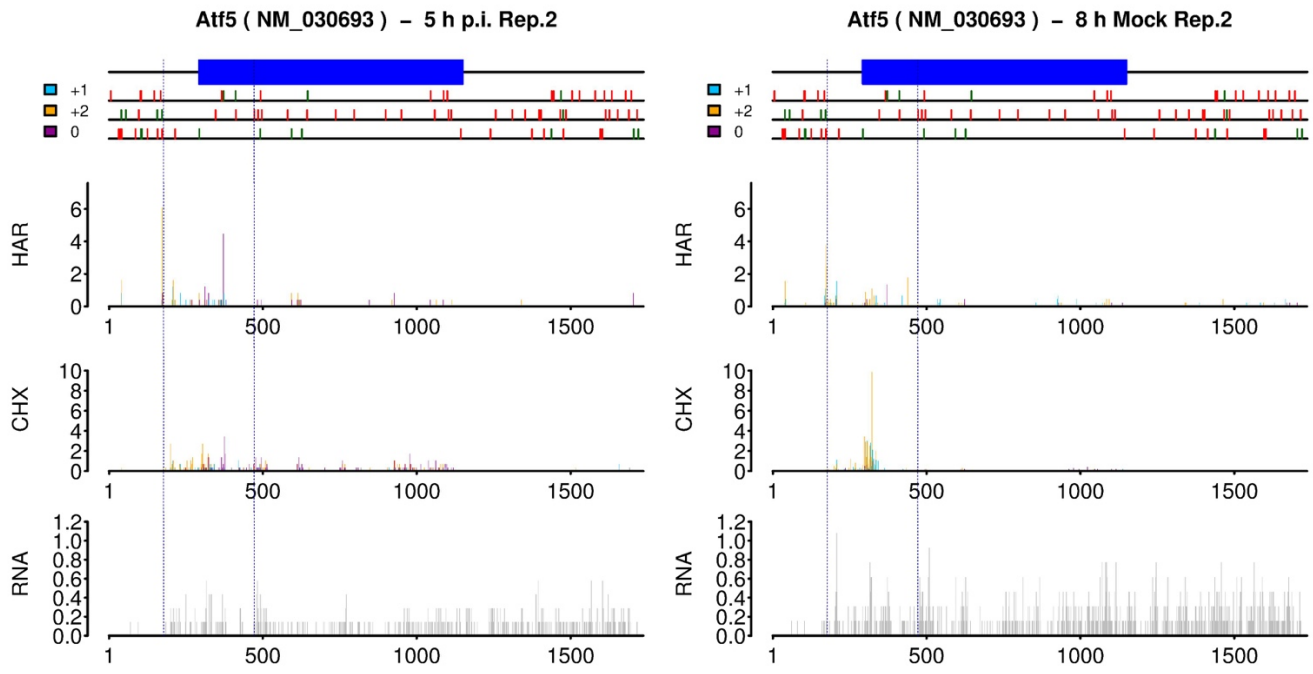


Figure 3.9 Atf5 RiboSeq plot. Atf5 RiboSeq and RNASeq read distributions compared between infected and uninfected cells, as for the Atf4 plot in Figure 3.8.

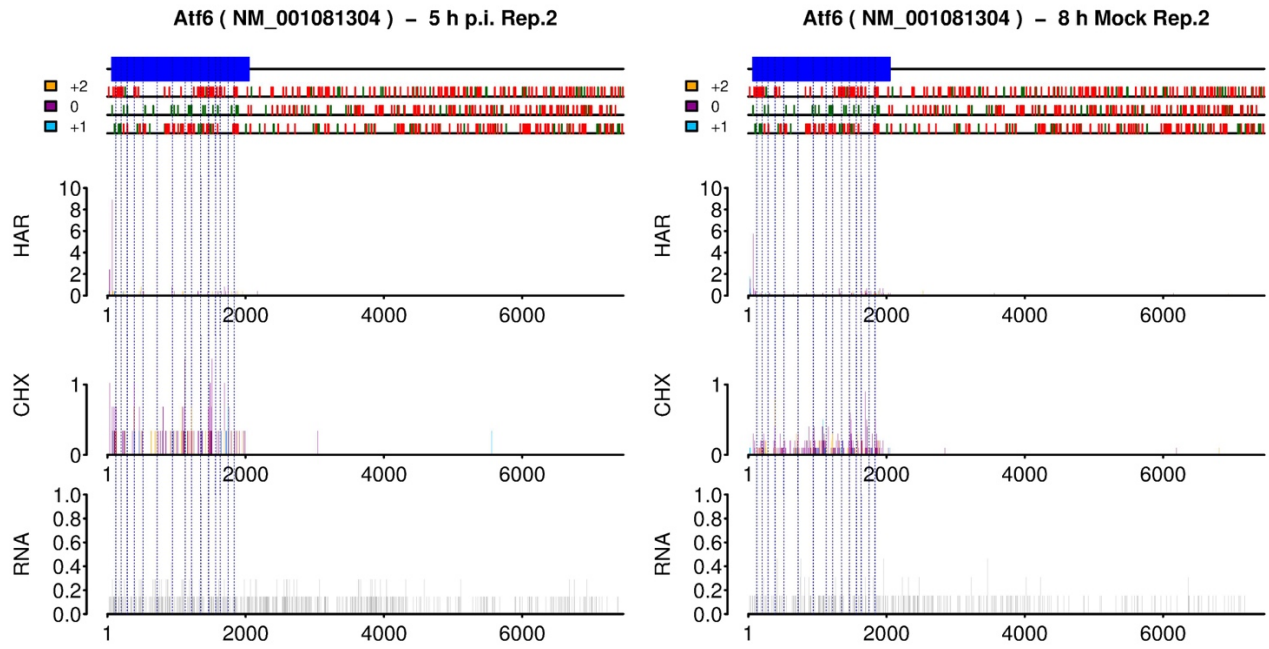


Figure 3.10 Atf6 RiboSeq plot. Atf6 RiboSeq and RNASeq read distributions compared between infected and uninfected cells, as for Atf4 plot in Figure 3.8.

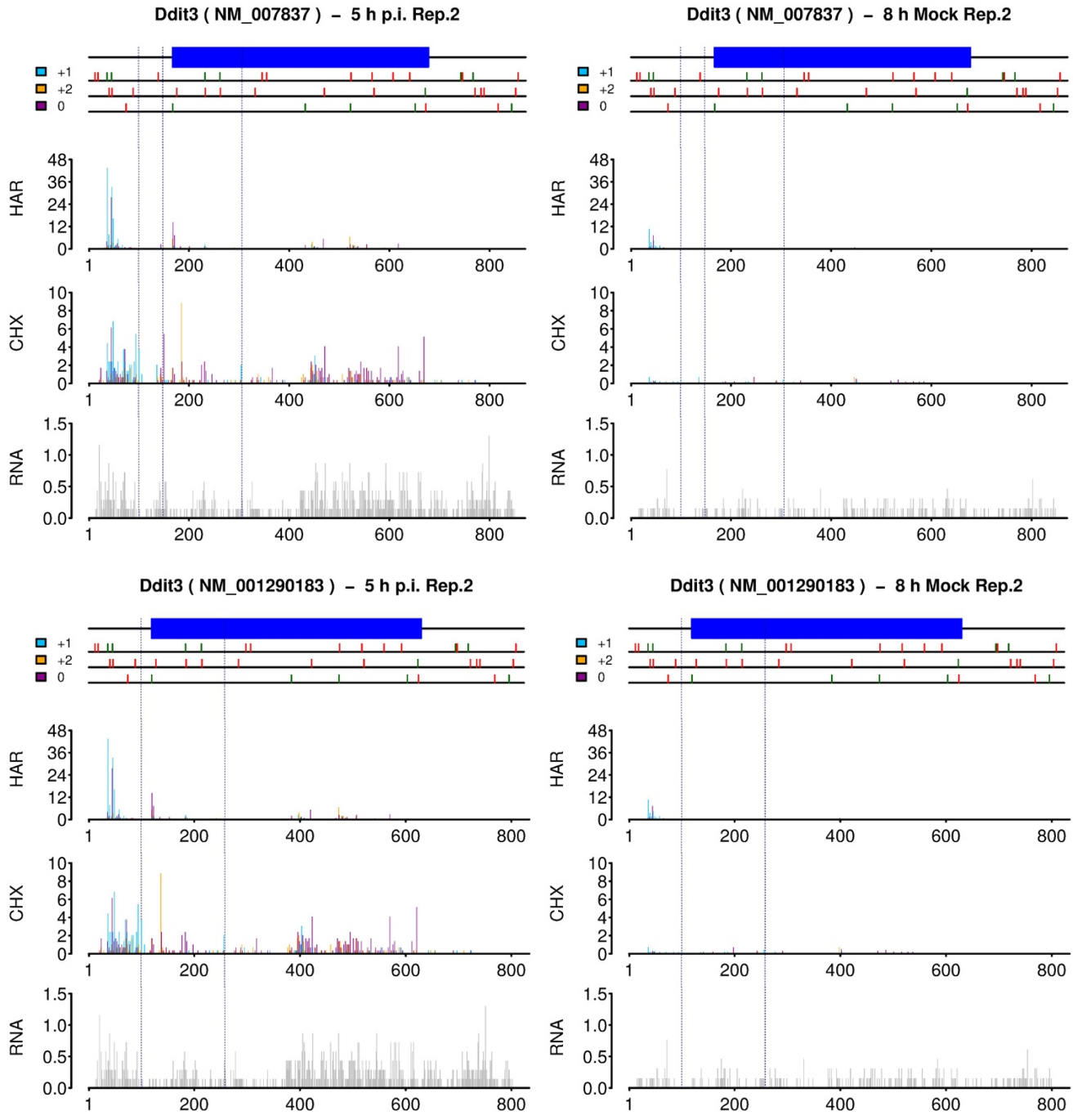


Figure 3.11 Ddit3/CHOP RiboSeq plot. Ddit3/Chop RiboSeq and RNASeq read distributions compared between infected and uninfected cells, as for Atf4 plot in Figure 3.8. Note that the increase in translation of the main ORF after MHV infection occurs despite an increase in initiation on the uORF AUG. Top panel shows data for one isoform of Ddit3 (NM_007837) while the bottom panel shows read alignments for the other annotated isoform of the gene (NM_001290183). We searched the UCSC EST database to see if we could find evidence for a third, unannotated isoform that would better explain the read distributions seen in these plots, but did not find strong evidence for a novel transcript.

that ribosomes may be scanning past the main initiation codon and initiating in the second of the ORF. For the UPR transcription factor Xbp1, an increase in the number of reads in the +2 reading frame was seen after the annotated stop codon, consistent with translation of the frameshifted Xbp1-s transcript that is produced by IRE-mediated excision of a 26-nt from the transcript (

Figure 3.12).

In addition to the various UPR regulators, we also observed two other miscellaneous genes that were differentially expressed and exhibited unusual translation patterns. In the case of the selenoprotein, Sepw1, virus infection led to an increase in the total number of read alignments and the potential synthesis of an N-terminally extended peptide (

Figure 3.13). Though we did see a non-significant increase in translation of the *Slc35a4* ORF during later time points in infection, the vast majority of RPFs mapped to the 102-codon uORF (

Figure 3.14). *Slc35a4* has been previously shown to be resistant to eIF2α-induced translational attenuation and its 102-codon uORF has been shown to be highly conserved (Andreev et al., 2015a). While we could not evaluate the resistance to translation shutoff in this transcript due to the low baseline levels of translation, the presence of reads in the uORF raises the possibility that the main ORF is translationally silenced by the upstream element. The role of these genes in the context of antiviral responses or viral replication is not known.

3.2.5 5′ RPF Loading Ratio

Previous meta-analyses of RPFs mapping to host mRNAs during MHV infection indicated that, as the infection progresses, an increasing proportion of reads map to the first codons of CDSs (

Figure 3.15; Irigoyen et al., 2016). Heat shock, proteotoxic chemicals and oxidative stress have been similarly shown to inhibit elongation and cause an accumulation of ribosomes in the 5′ portion of ORFs (Gerashchenko et al., 2012; Liu et al., 2013; Shalgi et al., 2013; Liu and Qian, 2014). The interaction of chaperones and the UPR-related effect on translation likely occurs as a nascent peptide starts to exit the ribosome, hence the arrest being in the first 30-50 codons. It has been noted that mRNAs susceptible to such elongation inhibition tended to encode N termini with stronger Hsp70 interacting motifs and that the extent of pausing

correlates with the hydrophobicity of the N termini (Liu et al., 2013a). Similarly to these previous studies, we used a relative loading ratio statistic (r5′ LR) to quantify elongation arrest by comparing the density of reads in 5′ end of each ORF with the density of RPFs in the remaining portion of the coding sequence in mock and infected replicates. Specifically, we divided the density of RPFs mapping to nucleotide positions 16-90 in each CDS by the density

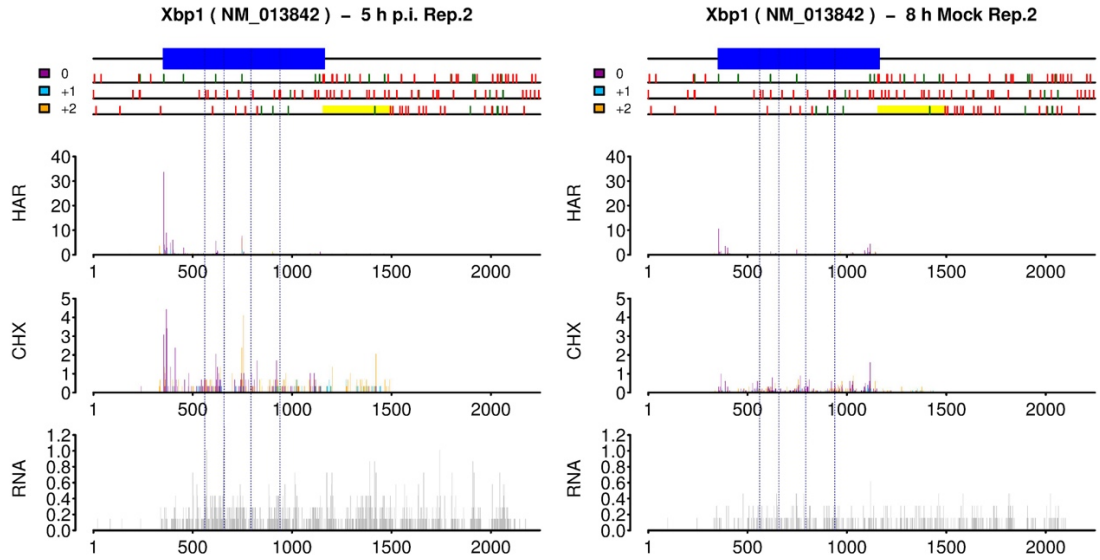


Figure 3.12 Xbp1-s RiboSeq plot. Xbp1 RiboSeq and RNASeq read distributions compared between infected and uninfected cells, as for Atf4 plot of Figure 3.8. Left plot panel is from a 5 h p.i. replicate, and the rightmost panel is from an 8 h mock replicate. Note that presence of reads in the +2 frame downstream of the annotated stop codon (yellow peaks) is derived from translation of the Xbp1-s frameshifted isoform.

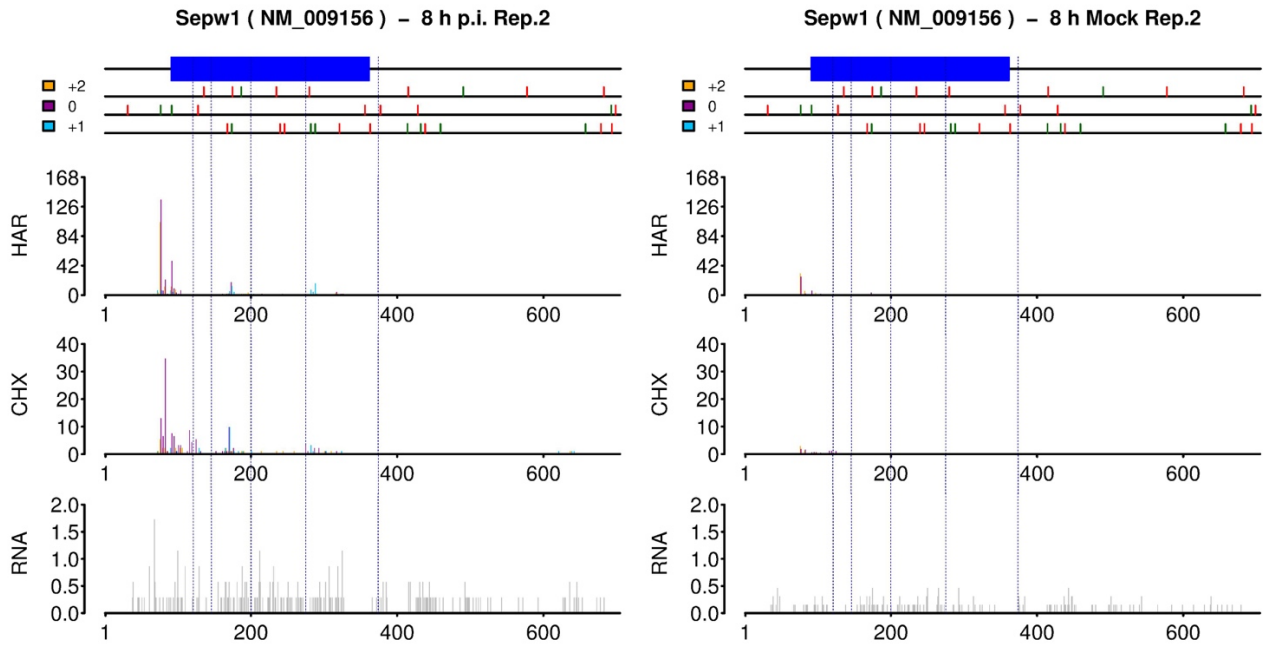


Figure 3.13 Sepw1 RiboSeq plot. Sepw1 translation increases during MHV infection and we observe reads aligning to an AUG upstream of the annotated start codon. Note that the red tick mark at position 128 in the 0-reading frame is a UGA codon that is used to encode a selenocysteine; the annotated stop codon is at position 358.

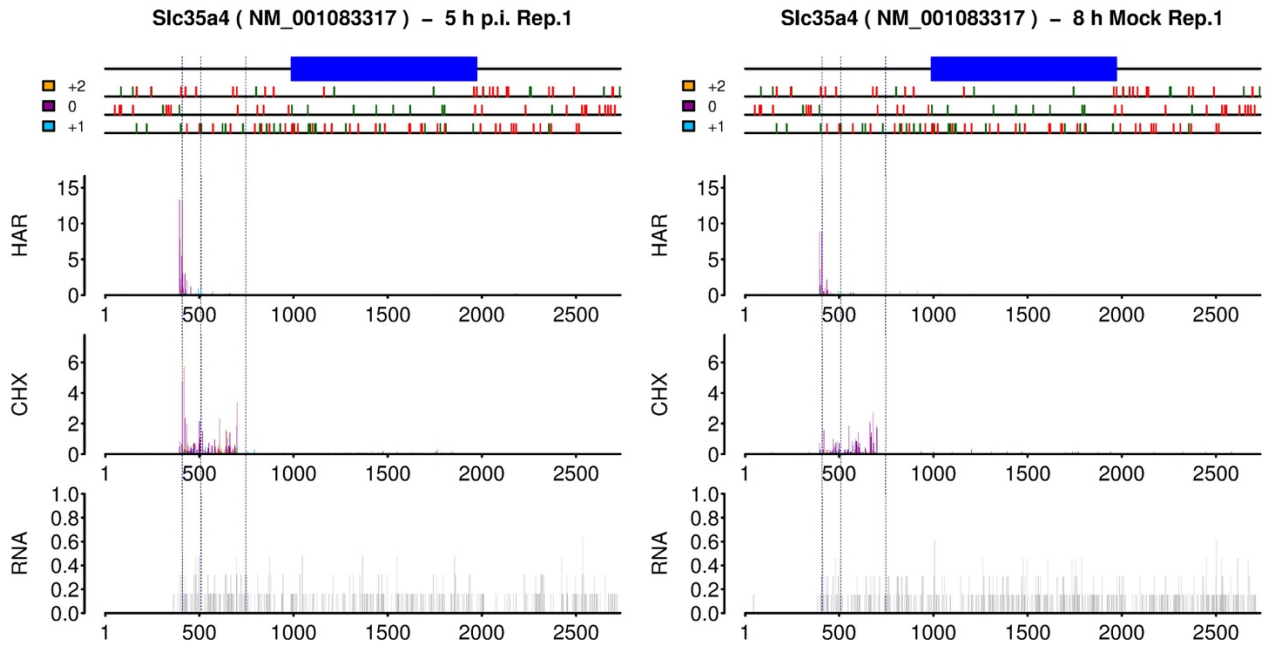


Figure 3.14 Slc35a4 RiboSeq plot. Slc35a4 RiboSeq and RNASeq read distributions compared between infected and uninfected cells, as for Atf4 plot in Figure 3.8. The 102-codon uORF is translated with excellent framing; the annotated ORF exhibits little translational activity in either mock or infected cells.

of RPFs mapping to all nucleotide positions from 91 until the annotated stop codon to a loading ratio for a given transcript in a given replicate. We subsequently divided the loading ratio of the infected replicate by the loading ratio of the uninfected replicate to generate the relative loading ratio. Genes with fewer than 10 reads mapping in any of the defined windows were discarded. The resulting loading ratios indicated that ~90% of mRNAs had an increase loading ratio due to viral infection (

Figure 3.16), suggesting some elongation inhibition. The number of transcripts with elevated loading ratios increased as the viral infection progressed from 5 h p.i. to 8 h p.i. For viral transcripts, we could not perform an analysis of loading ratios in infected versus uninfected cells, so we chose to analyse the transcripts at specific timepoints, with the loading ratios for viral transcripts compared to the distribution of cellular loading ratios (bottom left two panels in

Figure 3.16). This comparison indicated that MHV transcripts are not generally resistant to elongation inhibition. The relative loading ratio analysis did reveal that mitochondrial transcripts were generally recalcitrant to elongation inhibition, as would be expected from their physical sequestration and independent translation apparatus. As mentioned in Chapter 1, a previous study conducted in yeast suggested that some observations of RPF accumulation at the beginning of ORFs are due to cycloheximide treatment, which may offer an alternative or partial explanation for mitochondrial transcripts being resistant to perceived elongation inhibition in this study (Gerashchenko and Gladyshev, 2014).

A shortlist of cellular mRNAs that had a r5′ LR equal to, or less than one (suggesting decreased elongation inhibition due to virus infection) was compiled and analysed for possible commonalities to suggest resistance to the inhibition effect. There did not appear to be any correlation between loading ratios and localization to the mitochondrion or endoplasmic reticulum, or presence of a signal peptide. Similarly, the predicted biophysical properties of the nascent proteins (hydrophobicity, isoelectric point, proline content, aromatic amino acid content, etc.) did not appear to correlate with the loading ratios, in contrast to the relationship between hydrophobicity and pausing described by Liu et al., (2013). It is possible that the transcript-specific effect of MHV-induced elongation inhibition was not strong enough to allow for clear separation of signal from noise with this dataset. There may also be other relevant peptide characteristics that we did not identify and correlate with the loading ratios. Although some mRNAs showed resistance to elongation inhibition, potentially mediated by an RNA secondary structure or *trans*-acting factor, we did not observe any unifying features that could account for such resistance.

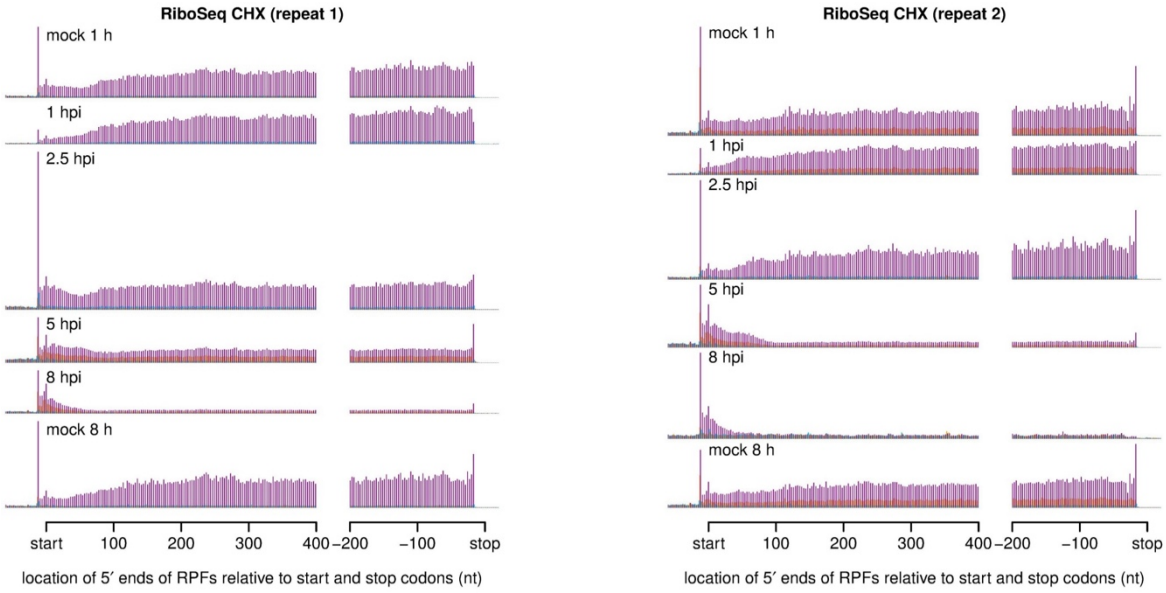


Figure 3.15 RPFs mapping relative to start and stop codons after MHV infection. Histograms of RPF 5′ end positions relative to annotated initiation and termination codons summed over all host RefSeq mRNAs for RiboSeq libraries from MHV-infected cells. Note that repeat 2 at 5 h p.i. seems more advanced than repeat 1 (see Figure 3.17 for comparison). Figure adapted from Irigoyen et al., (2016).

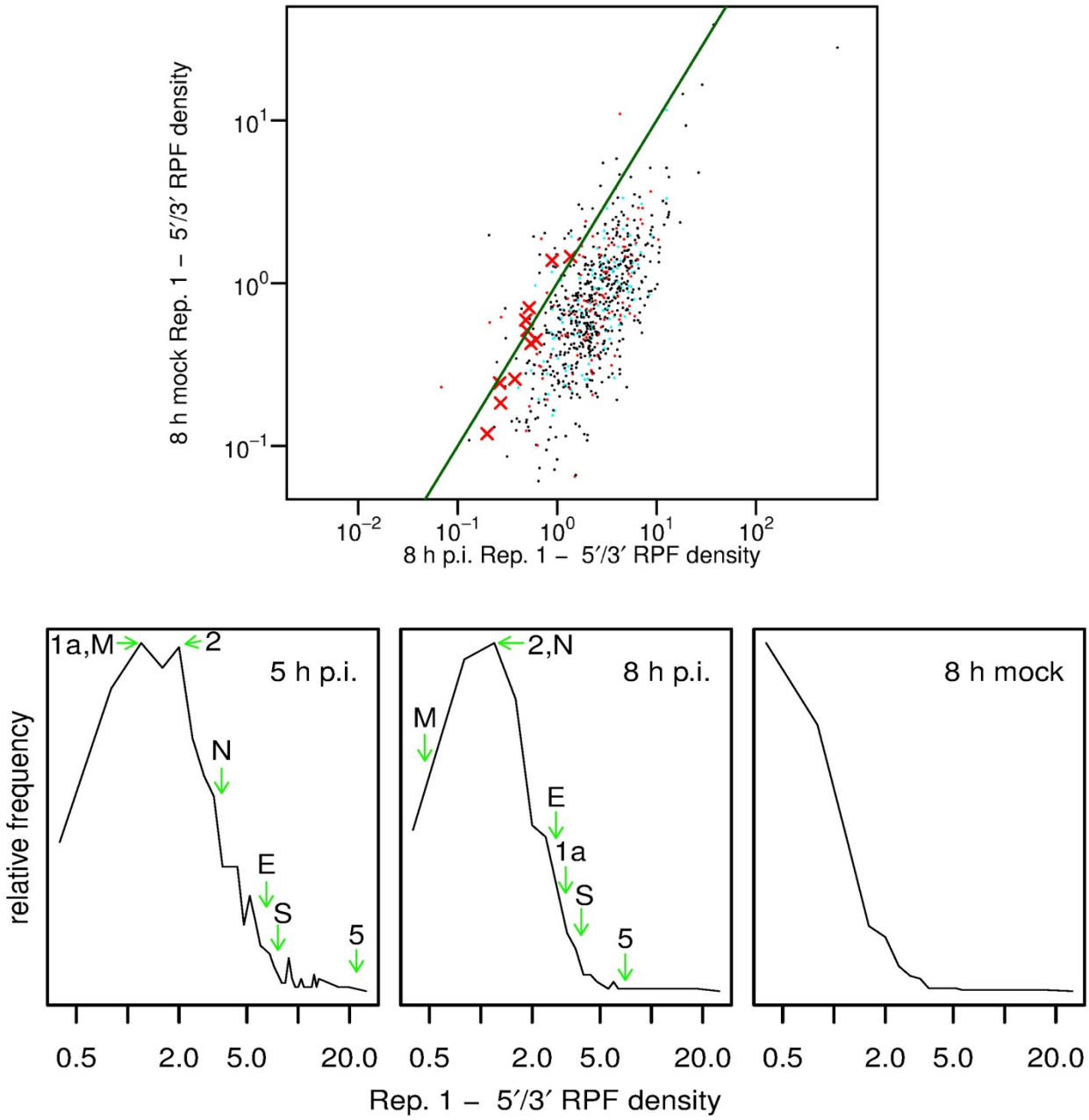


Figure 3.16 MHV induces elongation inhibition. Top panel shows a scatterplot of 5′/3′ RPF loading ratios under infected (8 h p.i.) and mock conditions. Only genes with more than 49 RPFs mapping between positions 16 and 90 nt in the 8 h mock are shown. CDSs predicted to encode mitochondria-targeted and secretome proteins are indicated with cyan and red points, respectively. Red crosses indicate mitochondrial transcripts such as *Atp8*. Bottom panels show histograms of relative loading ratios for 5 h p.i., 8 h p.i. and 8 hr mock (left to right, respectively), with positions of viral transcript loading ratios marked with green arrows. Data are shown for repeat one.

3.3 Discussion

This study represents the first high-throughput analysis of changes in a cellular translatome under coronavirus-induced stress. The results corroborated (and in some cases, qualified) several previous studies on coronavirus infection, the UPR and eIF2α phosphorylation-resistant translation. We dissected the activation of Xbp-1, Atf4 and Atf6, performed a meta-analysis of elongation inhibition, and found translational apparatus and transcript sequestration genes, but not interferon-stimulated genes, to be upregulated. The RiboSeq data provided additional evidence to support the hypothesis that one of the mammalian solute carrier protein transcripts is bicistronic.

Large scale elongational arrest occurred in many different mRNA species, consistent with previous studies on the roles of various MHV proteins as well as the effects of UPR induction (Raaben et al., 2007; Bechill et al., 2008; Fung et al., 2016). We observed changes in RiboSeq alignments that are consistent with uORF-mediated resistance to eIF2α-regulated translational attenuation in several transcripts, including those which were previously implicated as effectors of the UPR. While Versteeg et al. (2007) and Bechill et al. (2008) failed to detect Xbp1-s proteins in MHV-infected cells, which they attributed to translational shut-off, our data indicate that ribosomes are found in the frameshifted portion of the *Xbp1* ORF. It is possible that coronavirus infection leads to rapid degradation of this transcription factor or that the previous studies did not use sufficiently sensitive antibodies to detect the protein. Differences in cell lines or infection kinetics used may also possibly explain these discrepancies. Similarly, previous studies have noted an upregulation of Atf4 during MHV infection but have failed to detect induction of its downstream effectors, GADD34 and Ddit3 (Fung et al., 2016). However, we found evidence for an increase in translational efficiency of these transcripts at late time points post infection. While ribosome profiling provides data about translational activity, the rate of protein synthesis is only one determinant of protein abundance, which is also influenced by the rate of protein degradation. Additional studies with careful controls and different modalities will be needed to establish the influence of MHV infection on these components of the UPR.

The role of the Atf6 branch of the UPR in the context of MHV infection has been previously unresolved. While one study observed activation of pre-existing pools of Atf6 during MHV infection, it noted a decrease in total Atf6 levels as the infection progressed, which was attributed to translational arrest (Bechill et al., 2008). We did not detect significant changes in Atf6 expression in our Xtail analysis, though the transcript did not have as many RPFs mapping

to it as Atf4 under any condition. The Atf6 transcript has a 5′ UTR that is only 117 nucleotides long and lacks any potential CUG/AUG initiated uORFs, unlike Atf4 and Atf5. Our results therefore support the view that Atf6 expression does not increase under conditions of MHV-induced ER stress.

Previous authors have speculated on how MHV proteins are synthesized in the face of UPR-induced translational arrest (Fung and Liu, 2014). Our loading ratio analysis indicates that MHV transcripts are not systematically resistant to host translational shut-off; rather, it seems that the over-abundance of MHV transcripts makes them stochastically more likely to be translated during later-times points when competing with cellular transcripts for a limited pool of translation components. The UPR-mediated eIF2α phosphorylation may be favourable to viral replication in that it prevents translation of various factors involved in the anti-viral response. Indeed, experimental inhibition of the UPR response during MHV infection can reduce MHV infectivity (Dr Nerea Irigoyen*; personal communication*). One interpretation of this study is that MHV engages in a delicate balancing act; the UPR pathway is beneficial to it in that it upregulates phosphorylation of eIF2α, which reduces translation of immune mediating proteins and anti-viral factors; but at the same time, over-stimulation of the pathway can induce apoptosis and prevent release of progeny virions.

We noted changes in the expression levels of several components of the translational apparatus, which may be due to a cellular effort to restore homeostasis and compensate for decreases in translational competence during MHV infection. MHV infection has been shown to induce RNase L-independent cleavage of 28S rRNA and could be responsible for inducing these changes (Banerjee et al., 2002). We could not quantify the modulation of rRNA abundance in our datasets as our protocol is not designed for rRNA analysis (e.g. due to the use of RiboZero and DSN rRNA depletion techniques).

Coronaviruses have been noted to be poor inducers of type I IFN in cell culture (Spiegel et al., 2005; Roth-Cross et al., 2008; Fung and Liu, 2014), though this may not be the case *in vivo* (Ireland et al., 2008). Our differential expression results corroborate this interpretation – interferon genes were not found to be significantly upregulated at either the translational or transcriptional level in response to MHV infection at 5 h p.i. or 8 h p.i. One limitation of our study is that we only had two infected replicates per time point and that at 5 h p.i. there appeared to be a significant difference in infection progression - the second replicate was more advanced in infection (Figure 3.17; Irigoyen et al., 2016). This contributed to a larger variance in read counts between the two infected replicates at this time point, and may have reduced sensitivity for differentially expressed genes.

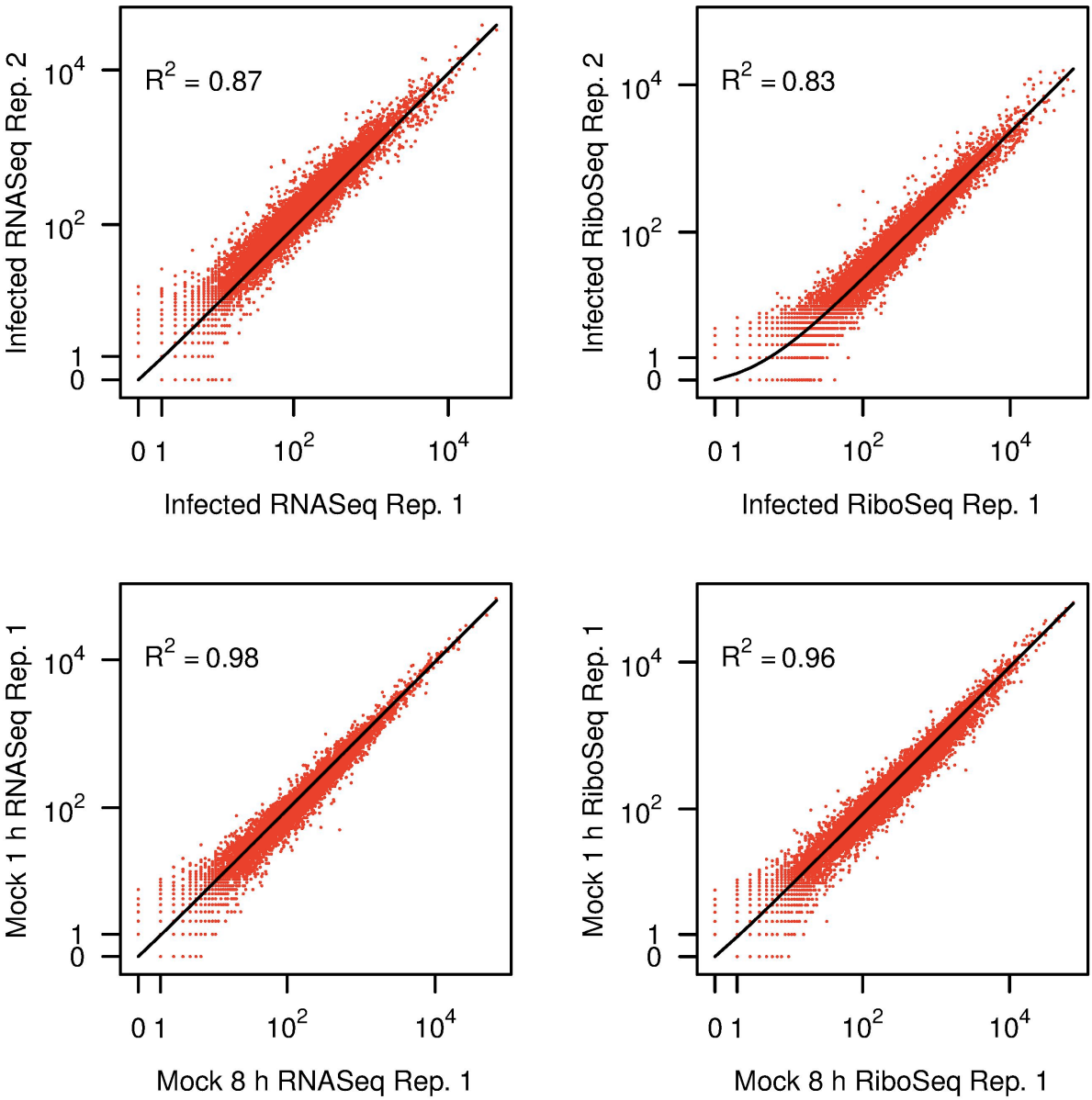


Figure 3.17 Inter-replicate consistency. Top left and right panels show consistency between gene read counts for 5 h p.i. RNASeq and RiboSeq samples respectively. There is some variation in reads counts in the first replicate relative to the second, suggesting differences in infection progression. Bottom left and right panels show consistency between RNASeq and RiboSeq (respectively) gene read counts for 1 and 8 h uninfected controls.

TIA was translationally upregulated - consistent with a previous study that noted the formation of stress granules due to either MHV infection or sodium arsenite treatment in LR7 cells (Raaben et al., 2007). An important question is whether the sequestered mRNAs in SGs/P-bodies can interfere with ribosome profiling analyses. SGs that (partially) sequester a transcript species may lead to the false calling of differential translational efficiency. Future studies will be needed to determine if transcripts localised to these cytoplasmic structures mix with other nucleic acid fragments during library preparation.

Slc35a4 was not significantly increased in our Xtail differential expression analysis, though we did observe a very large number of RPFs mapping in all replicates to the Slc35a4 102-codon uORF. This result provides another argument in support of this uORF encoding a functional peptide. Previous ribosome profiling noted RPFs mapping to this uORF and phylogenetic analyses showed a high degree of evolutionary conservation (Andreev et al., 2015a). Two mass spectrometry studies also contained data consistent with expression of this peptide (Vanderperre et al., 2013; Kim et al., 2014). The increasing number of observations suggests that Slc35a4 may indeed be a rare example of a mammalian bicistronic mRNA and warrant further investigation.

In conclusion, this study provides a survey of coronavirus effects on the cellular translatome and complements multiple previous investigations of the UPR and MHV infection. During the course of work, I helped improved one of the few specialised ribosome profiling differential expression analysis tools (Xtail) and developed custom methods for visualising RiboSeq distributions which are particularly useful in the case of genes with multiple or putative isoforms (discussed in section 3.2.4). These tools will be made publicly available via a GitHub repository. Future work may serve to integrate some functionalities of these tools with other RiboSeq visualization tools (Thorvaldsdóttir et al., 2013; Michel et al., 2014; Chung et al., 2015). We hope that the results of our analyses will help inform further investigations of coronavirus–UPR interactions and may help identify new targets for antiviral agents.

Chapter Four

Computational Analyses of Flavivirus Ribosome Profiling Data

4 CHAPTER FOUR: COMPUTATIONAL ANALYSES OF FLAVIVIRUS RIBOSOME PROFILING DATA

4.1 Introduction

The *Flaviviridae* have been the subject of intense medical study for over a century; yellow fever and dengue fever were, respectively, the first and second human diseases shown to be caused by a viral agent (Lindenbach et al., 2013). These pathogens have been implicated in numerous historical epidemics, and their importance was highlighted in the award of the 1951 Nobel prize to Max Thieler for developing the first yellow fever vaccine. Zika virus, first isolated in 1947, has been the subject of increasing scientific focus due to a major recent outbreak that has led to ~1.3 million new infections and reports of birth defects (Baden et al., 2016; Plourde and Bloch, 2016). However, despite intense scrutiny from the biomedical community, this group of viruses continues to cause hundreds of thousands of deaths globally each year.

Previous reports have highlighted the importance of efficient genome translation as a primary determinant of flavivirus infectivity (Edgil et al., 2003; Pierson and Diamond, 2013). We have used ribosome profiling to survey the dengue virus (DENV) translatome in *Aedes albopictus* and human cell lines to further investigate several unresolved points of the viral replication cycle.

Several studies have shown that the DENV genome is replicated and utilised in a temporally sensitive manner (Alvarez et al., 2008; Paranjape and Harris, 2009a; Alcaraz-Estrada et al., 2010; Gebhard et al., 2011; Viktorovskaya et al., 2016). DENV does not appear to shut-off host translation, so there are unresolved questions about the mechanisms that enable it to co-opt the cellular protein synthesis apparatus (Villordo et al., 2010). It has been previously debated whether the viral structural and non-structural regions of the polyprotein ORF are translated with equal efficiency. Japanese encephalitis virus (JEV) and West Nile virus (WNV) utilize efficient (30-50%) programmed ribosomal frameshifting in their NS2A regions to encode produce the NS1' protein (Mason, 1989; Chen et al., 1996; Firth and Atkins, 2009; Melian et al., 2010, 2014; Takamatsu et al., 2014). This frameshift signal also downregulates expression of the viral RNA-dependent RNA polymerase, and we were interested in seeing if a similar variation in translation levels occurs in DENV (Ahlquist, 2006). To this end, we sought to investigate the kinetics of DENV positive and negative sense transcript production as well as translation over the course of an infection cycle. DENV RNAs are known to interact with numerous cellular RBPs and the polyprotein contains numerous transmembrane domains (Miller et al., 2007; Oostra et al., 2007; Phillips et al., 2016). Thus we sought also to identify

DENV translational pause sites that may further inform its molecular biology, as well as how these may vary given the different constraints imposed by the mosquito and human intracellular environments. The canonical DENV ORF begins at an AUG that is downstream of several other AUG sites with equally 'valid' initiation contexts and is recognized due to a RNA structure-programmed pausing event (Clyde and Harris, 2006). We sought to verify if there is potential for alternative translation products to be made via recognition of one of these upstream AUGs. DENV has also been shown to produce a truncated version of its genome known as the subgenomic flavivirus RNA (sfRNA), whose role has not been fully characterised (Chapman et al., 2014a). We were also interested in uncovering whether the sfRNA is actively produced and translated during DENV infection.

This chapter will detail the DENV genome and its replication cycle, as well as insights from previous studies that drove this investigation. Subsequently, various analyses aimed to gauge the suitability of using profiling for studying DENV replication will be discussed. Finally, a ribosomal profiling analysis of DENV infection in human and mosquito cells will be described.

4.1.1 Flaviviruses

The *Flaviviridae* are a family of enveloped viruses with positive-sense, monopartite single-stranded RNA genomes that infect various mammals and spread through arthropod vectors (some are insect-specific). The prototypical flavivirus, Yellow Fever virus (YF) is known for its ability to induce jaundice in patients, hence the taxon name – *flavus* meaning 'yellow' in Latin (Lindenbach et al., 2013). There are currently four recognized genera in the family – *Pestivirus*, *Pegivirus*, *Hepacivirus* and *Flavivirus*. As mentioned before, these viruses are of high clinical relevance (Table 4.1).

The *Flaviviridae* have enveloped, spherical virions with a lipid bilayer, ~50 nm in diameter. The envelope (E) glycoproteins on the surface of the virions facilitate receptor binding and cell entry, and are the major antigenic determinant that helps inform classification. Members of the *Flavivirus* genus can be stratified by their arthropod vector – either tick or mosquito borne. The central flavivirus of this text (DENV), has been grouped into five distinct serotypes that exhibit up to 30% genome sequence divergence and which may undergo recombination (Holmes, 2009; Carvalho et al., 2010; Normile, 2013; Mustafa et al., 2015). The presence of distinct DENV strains has been one of the main barriers to the development of an effective DENV vaccine as, for the vaccine to be effective, it must protect against the various

***Flaviviridae* example**	***Flaviviridae* Genus**
Yellow fever virus (YFV)	*Flavivirus*
Dengue virus (DENV)	*Flavivirus*
Japanese encephalitis virus (JEV)	*Flavivirus*
West Nile virus (WNV)	*Flavivirus*
Zika virus (ZIKV)	*Flavivirus*
Kyasanur forest disease virus	*Flavivirus*
Tick-borne encephalitis virus (TBEV)	*Flavivirus*
Murray Valley encephalitis virus (MVEV)	*Flavivirus*
Hepatitis C virus (HCV)	*Hepacivirus*

Table 4.1 Clinically relevant *Flaviviridae* . The *Flaviviridae* include several well-known and emerging pathogens that have spread from zoonotic sources to humans.

serotypes of the virus quasi-species (Normile, 2013). Vaccination against one serotype leads to production of antibodies that are cross-reactive against but do not neutralize the other serotypes (Whitehead et al., 2007; De Alwis et al., 2014). These antibodies facilitate viral replication upon new infection by facilitating entry of active DENV into macrophages, a process termed antibody-dependent enhancement. The clinical effect of this reaction is dengue haemorrhagic shock. An extensive effort to create a multi-valent vaccine failed in clinical trials in 2012 as it did not protect against a strain of the virus that was prevalent in Thailand at the time (Sabchareon et al., 2012; Schwartz et al., 2015).

4.1.2 Dengue virus genome and proteins

The DENV genome consists of a non-segmented ~10.7 kb RNA molecule that has a single ORF that is translated as a single polyprotein (

Figure 4.1). As in all single-stranded, positive-sense RNA viruses, the DENV genome has three functions – as the genetic material that is distributed within progeny virions, as a template during RNA-directed RNA replication, and as the mRNA for translation of the polyprotein (Pierson and Diamond, 2013). The RNA molecule appears to be translated in a similar (cap-dependent) fashion to cellular transcripts and has a 5′ m^7GpppAmN cap, though a recent report has described an alternative, cap-independent translation mechanism for the virus that utilizes the stem-loop B (Dong et al., 2007; Paranjape and Harris, 2009; Pérez Dominguez et al., 2015). Unlike cellular transcripts, the DENV RNA has no poly-A tail, though its 3′ UTR does enhance translation (Holden and Harris, 2004; Chiu et al., 2005).

The DENV 5′ and 3′ ends contain several conserved RNA sequences and structures that facilitate replication and have the capacity to interact and circularise the genome (Figure 4.2). The 5′ ends contains the motifs 5′ conserved sequence (5′CS), 5′ upstream AUG region (5′UAR), 5′ stem-loop (5′SL) and capsid hairpin (cHP). The 5′UAR can fold into a RNA secondary structure as well, with the polyprotein AUG located at the 3′ end of this region. The cHP, located in the capsid encoding portion of the ORF, functions in a position-dependent (sequence-independent) manner to induce ribosome stalling over the ‘main’ ORF initiation codon to ensure initiation occurs at this site rather than at one of the potential upstream AUG sites (Clyde and Harris, 2006). The 3′ UTR contains the 3′ stem loop (3′SL), 3′ conserved cyclisation sequences (3′CS1), 3′ UAR, as well as two 3′ pseudoknots (PK1, PK2) (Kieft et al., 2015; Villordo et al., 2015). The ~120 nt 3′SL binds to NS5, NS3 and NS2A and elongation factor 1A (EF1A) as well as polypyrimidine tract binding protein (PTB), though the purpose of

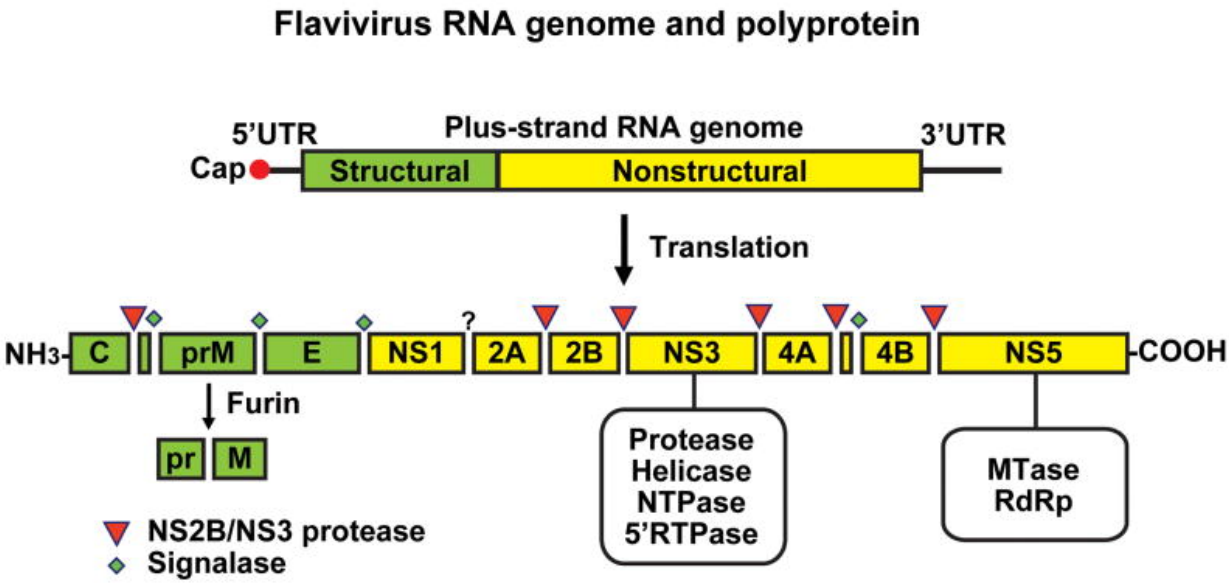


Figure 4.1 Dengue virus genome. Flavivirus genomes are roughly 10.8 kb positive-sense RNA molecules and include one large ORF that encodes a polyprotein. The N-terminal portion of the polyprotein contains structural proteins and the remainder contains replicase proteins. Figure adapted from Sampath and Padmanabhan (2009).

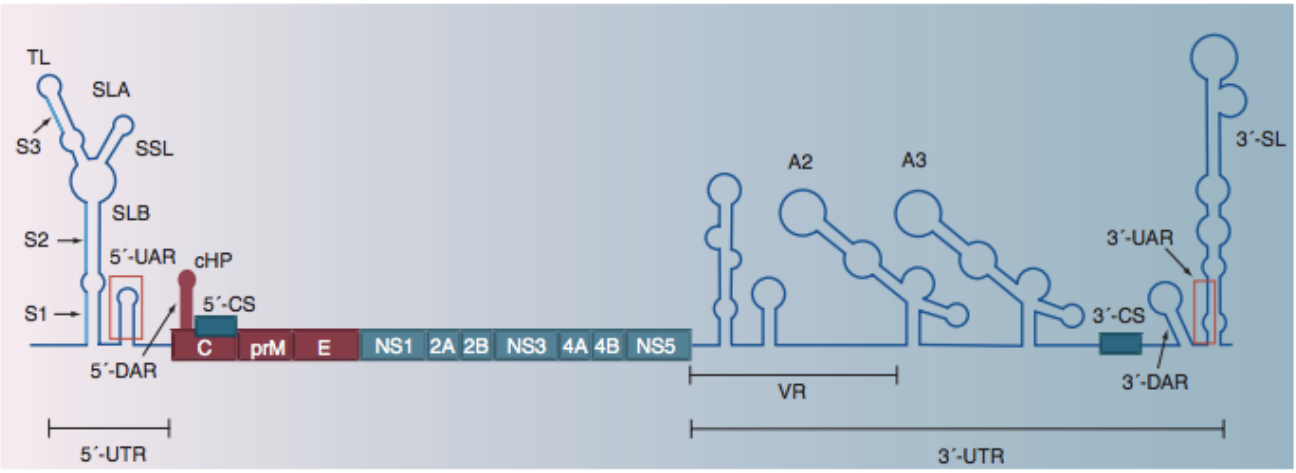


Figure 4.2 Secondary structures in the 5′ and 3′ DENV RNA UTRs. Various conserved sequences and secondary structures in the DENV RNA 5′ and 3′ UTRs, including several that are important for initiation of replication or translation. Figure adapted from Alcaraz-Estrada et al. (2010).

this last interaction is currently unknown (De Nova-Ocampo et al., 2002; Paranjape and Harris, 2009). Functional studies have shown that the 3′SL enhances translation of DENV RNA and antisense oligos against the 3′SL inhibit translation (Holden and Harris, 2004; Chiu et al., 2005; Holden et al., 2006). Certain point mutations in the 3′SL can abrogate DENV infectivity in mosquito cells without affecting reproduction in human cells (Villordo and Gamarnik, 2013), illustrating the notion that some components of the DENV genome may be engineered specifically for replication in one of its two hosts.

The DENV RNA molecule can shift between various spatial conformations that enact various structure-function relationships (Alvarez et al., 2008; Villordo and Gamarnik, 2009). The DENV 5′/3′ UAR and 5′/3′ CS regions can undergo long-range base pairing, causing the RNA molecule to circularise. This conformational change brings the 3′ site of minus-strand replication initiation into proximity with the 5′ SL-bound NS5 (Alvarez et al., 2008; Pierson and Diamond, 2013). During the course of this circularization, several secondary structures in the 5′ and 3′ UTRs are destabilised and the translation initiation site becomes occluded.

The functional DENV proteome consists of three structural proteins – envelope (E), membrane (M) and capsid (C), and seven non-structural proteins – NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5 (
Figure 4.1). The structural proteins are located in the N-terminal portion of the polyprotein while the non-structural proteins are derived from the C-terminal portion. The M protein is formed when the precursor M (prM) peptide is cleaved by cellular furin in the low-pH environment of the *trans*-Golgi apparatus (Yu et al., 2008). The C/prM, prM/E and E/NS1 links are cleaved by cellular signal peptidase, while NS2A/NS2B, NS2B/NS3, NS3/NS4A, and NS4B/NS5 are separated by the NS2B serine protease. The NS1/NS2A junction is cleaved by an ER-resident cellular protease whose identity is presently unresolved (Falgout and Markoff, 1995; Pierson and Diamond, 2013).

Of the seven non-structural proteins, only two have been shown to have catalytic activity and the roles of some of the proteins are not completely resolved. NS5 is a ~103 kD phosphoprotein with several catalytic domains (Davidson, 2009). The C-terminal domain has RNA-dependent RNA polymerase activity while the N-terminal domains have methyltransferase and guanylyltransferase activity (Issur et al., 2009; Zhao et al., 2015; Galiano et al., 2016). The viral RNA cannot utilise cellular cytoplasmic or nuclear enzymes for these latter functions as the replicated DENV nucleic acid is compartmentalised in ER-membrane derived compartments. NS5 functions on the ER surface but contains a nuclear localization signal that enables its import into the nucleus. Compartmentalization of some of the NS5 protein

in the nucleus is thought to be a programmed inhibition designed to downregulate viral replicase activity during the infection cycle (Davidson, 2009; Rawlinson et al., 2009). NS3 is a ~70 kD protein with RNA triphosphatase, serine protease and RNA helicase domains. NS3 activity can contribute to formation of invaginations of the ER membrane (Matusan et al., 2001; Benarroch et al., 2004; Erbel et al., 2006; Aleshin et al., 2007). NS1 is a ~46 kD glycoprotein that is translocated into ER. Its role has not been fully characterised, but data indicates it anchors the viral replication complex to membranes of the ER (Noisakran et al., 2008; Gutsche et al., 2011). NS2B is a ~14 kD membrane-associated protein which helps anchor NS3 and serves as a cofactor for the NS3 serine protease (Bollati et al., 2010; Yang et al., 2011). NS2A is a ~22 kD hydrophobic protein that interacts with the 3′SL and can inhibit interferon signalling (Muñoz-Jordan et al., 2003; Pierson and Diamond, 2013; Dalrymple et al., 2015). NS4A is a ~16 kD hydrophobic protein that interacts with NS1 and helps induce invaginations of the ER membranes as well as to anchor the replication complex to the virus-induced membranes (Chua et al., 2004; Miller et al., 2007). NS4A also helps block interferon signalling (Muñoz-Jordan et al., 2003). NS4B is a ~27 kD hydrophobic protein that interacts with and modulates NS3 activity, causing NS3 dissociation from ssRNA and promoting dsRNA unwinding (Umareddy et al., 2006). The protein also helps block interferon signalling (Dalrymple et al., 2015).

Besides the numerous *cis*-acting RNA interactions and dengue protein/RNA interactions, multiple studies have investigated the interplay between cellular *trans*-acting factors and the viral genome (Alcaraz-Estrada et al., 2010). UV-induced crosslinking, mobility-shift assays and RNA-affinity chromatography have been used to identify cellular proteins that bind to *cis*-acting elements of flavivirus RNAs (De Nova-Ocampo et al., 2002; Yocupicio-Monroy et al., 2003; García-Montalvo et al., 2004; Paranjape and Harris, 2007; Emara and Brinton, 2007). These investigations identified elongation factor 1a (EF1a), polypyrimidine tract binding protein (PTB), La, T-cell intracellular antigen-1 (TIA-1), the related protein (TIAR), Y Box binding protein-1 (YB-1), calreticulin, PDI and hnRNP A1, A2/B1 and Q as interacting factors. As discussed later, the interactions of DENV RNA with some of these proteins may be to act as a ‘sponge’ to soak up factors that may otherwise prove deleterious to viral replication. PTB on the other hand is known to facilitate processing and conformational changes in transcripts, and may be recruited by the DENV mRNA in order to facilitate efficient translation in the absence of a poly-A tail.

More recently, Viktorovskaya et al. (2016) used thiouracil cross-linking mass spectrometry and a siRNA knockdown screen in DENV-infected cells to identify hnRNP F, embryonic stem cell-specific 5-hydroxymethylcytosine-binding protein (HMCES) and RNA-

binding motif protein X chromosome (RBMX) interactions in the early stages of viral life cycle. PTB and non-POU domain-containing octamer-binding protein (NONO) were found to bind the dengue RNA in later stages. Phillips et al. (2016) used UV cross-linking followed by antisense-mediated affinity purification and mass spectrometry to identify the RNA chaperone cold shock domain-containing protein E1 (CSDE1) and polyadenylate-binding protein (PABPC1) as DENV RBPs. Marceau et al. (2016) used a CRISPR knockout screen in mutagenized HAP1 cells infected with DENV-2 to analyse additional factors besides RBPs that are important for the life cycle. Their investigation identified ER-associated multi-protein complexes involved in signal sequence recognition, *N*-linked glycosylation and ER associated degradation. Knockout of MAGT1, STT3A and STT3B (three components of the oligosaccharyltransferase (OST) complex) was shown to almost completely inhibit DENV replication; these proteins were also shown to interact directly with viral non-structural proteins.

4.1.3 Subgenomic flavivirus RNA

DENV (and related flaviviruses) produce a ~0.4 kb subgenomic RNA that has been shown to contribute to cellular pathogenicity (Pijlman et al., 2008; Chapman et al., 2014a). function of the sfRNA is still unresolved, but it is thought to help repress the XRN1-mediated RNA degradation and small RNA interference pathways, as well as inhibiting the type-I interferon response (Clarke et al., 2015). Xrn1 is a highly efficient 5′-3′ exoribonuclease that mediates RNA decay in eukaryotes (Nagarajan et al., 2013). RNA structures in the DENV 3′ UTR stall Xrn1 during enzyme-mediated degradation of the DENV genome (

Figure 4.3), though it is not completely resolved which of the 3′ UTR structures mediates the stalling (Chapman et al., 2014a, 2014b; Clarke et al., 2015; Charley and Wilusz, 2016). The viral RDRP, which moves in the opposite direction (3′ to 5′) relative to the exonuclease is not hindered by the secondary structures. Pijlman et al. (2008) showed that the sfRNA is produced in both human and mosquito cells, consistent with the high levels of conservation in the RNA decay pathway. Multiple cellular proteins have been shown to interact with the sfRNA, which has caused some to propose that it acts as a "sponge" for cellular RBPs (Roby et al., 2014; Bidet and Garcia-Blanco, 2014; Bidet et al., 2014; Charley and Wilusz, 2014; Clarke et al., 2015). One study reported the binding of G3BP1, G3BP2 and CAPRIN1 to the sfRNA; these proteins are known to facilitate stress granule formation and their interactions with this nucleic acid may be the reason that SG formation is not seen in DENV infected cells (Bidet et al., 2014).

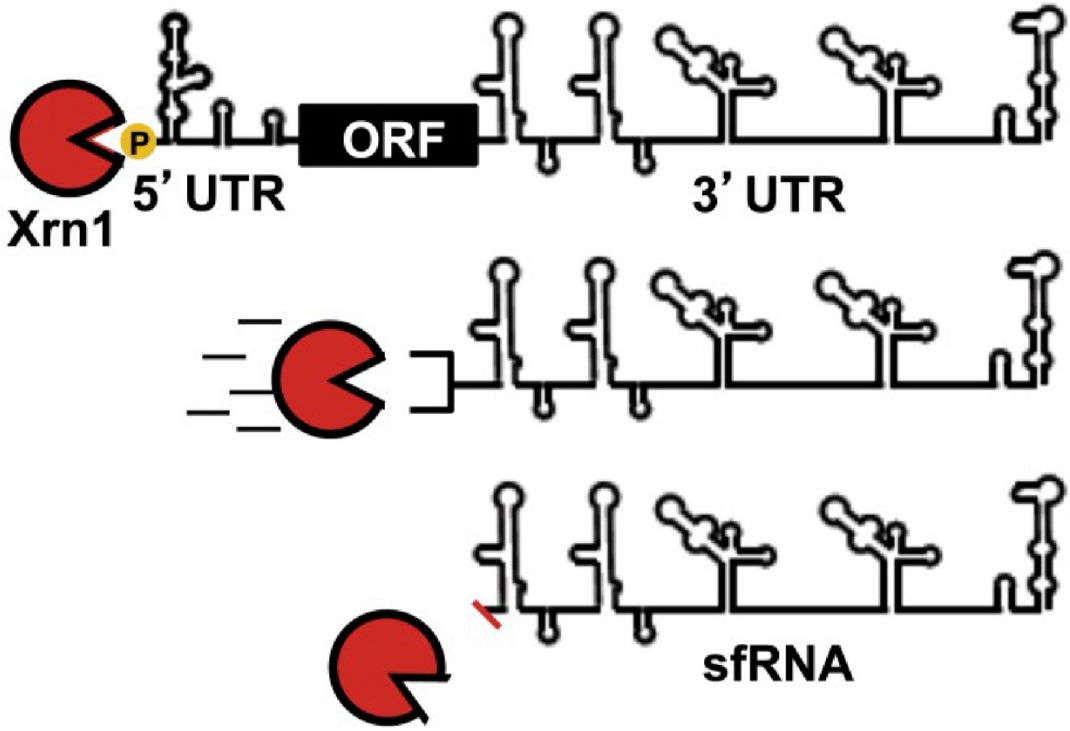


Figure 4.3 Generation of sfRNA by Xrn1 nuclease. The cellular Xrn1 5′-3′ exonuclease begins degradation of the DENV RNA from the 5′ UTR and processes until it encounters a pseudoknot in the 3′ UTR that causes it to stall and eventually dissociate. The resulting sfRNA is ~0.4 kb. Figure adapted from (Chapman et al., 2014b).

4.1.4 Dengue virus replication cycle

The DENV replication cycle (
Figure 4.4) begins with E glycoprotein interactions and receptor-mediated endocytosis. DENV receptors are not fully characterised, though it has been noted that the virus may utilise different cellular proteins for binding and entry into different tissue types and hosts (Pierson and Diamond, 2013). For dendritic cells, these include heparin sulphate and C-type lectin Dendritic cell-specific intercellular adhesion molecule 3-grabbing nonintegrin (DC-SIGN) (Navarro-Sanchez et al., 2003; Tassaneetrithep et al., 2003; Lozach et al., 2005; Artpradit et al., 2013). The mannose receptor and GRP78 have been identified as liver cell DENV receptors (Jindadamrongwech et al., 2004; Reyes-Del Valle et al., 2005; Cabrera-Hernandez et al., 2007; Miller et al., 2008). Other known DENV entry molecules include HSP70/HSP90 in macrophages and mosquito cells (Reyes-Del Valle et al., 2005; Das et al., 2009) as well as opsinizing immunoglobulin Fc receptors on various immune cells (Chotiwan et al., 2014).

Attached DENV particles diffuse across the cell surface until they encounter a clathrin-coated pit and become internalised into an endosome (van der Schaar et al., 2008). In an acidic endosome, E protein dimers on the virion shift into fusion-capable trimers and expose the fusion-peptide, which inserts into the endosomal membrane, allowing release of the viral RNA into the cytoplasm. The released dengue mRNA is transported to the ER and used to produce an initial set of viral proteins that can mediate replication. It is thought that at least a portion of the viral transcripts undergo host-factor mediated circularization prior to commencing translation (Paranjape and Harris, 2009). After the polyprotein is synthesized and cleaved by host and NS3 proteases into functional units, *cis*-acting structures in the DENV transcript (discussed earlier) and binding of NS5 with the 5′SL induces genome circularization and initiation of minus-strand RNA synthesis (Filomatori et al., 2006, 2011). Phosphorodiamidate morpholino oligos that target the 5′SL severely reduce viral translation (Kinney et al., 2005; Holden et al., 2006). The NS5 RDRP synthesizes a minus-strand RNA, which then is used as a template for production of plus-strand copies of the genome. Previous studies have found the (+)/(-) strand ratio to be 10:1, suggesting that each minus strand directs multiple rounds of copying (Pierson and Diamond, 2013). The synthesis of each new strand was measured to take 12-15 minutes (Paranjape and Harris, 2009).

NS5 processes the 5′ end of the emerging nucleic acids, and subsequently positive-sense transcripts are translated to make more DENV proteins as well as to be packaged into

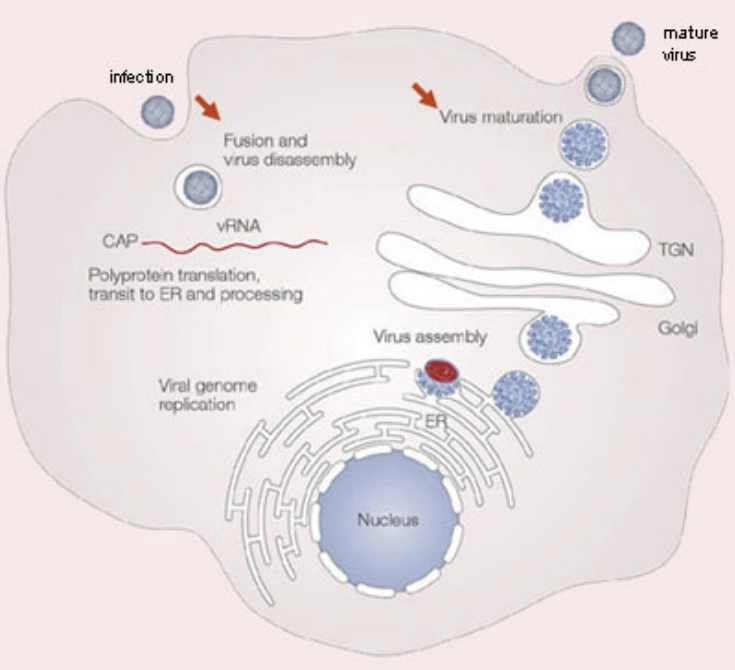


Figure 4.4 DENV replication cycle. DENV undergoes receptor-mediated endocytosis and pH-dependent membrane fusion. The virus induces invaginations of the ER membranes, forming compartments that act as foci for the various processes of virion replication and assembly. The progeny virions are released via the cellular secretory pathway. The NS5 polymerase encoded by DENV is highly error-prone (1 in 10^3-10^5), leading to high levels of virus variability. Figure adapted from *denguevirusnet.com*.

assembling virions. The vesicles containing the replication complexes are adjacent to forming DENV particles (Welsch et al., 2009), spatially coupling these processes.

The assembly process begins with C protein dimers pairing with viral RNA and budding into ER membrane spaces containing E-prM glycoprotein complexes. Transit of the immature virion through the acidified compartments of the trans-Golgi network induces conformational changes in E proteins. This pH-dependent conformational change also exposes prM molecules to cleavage by furin-like serine protease (van der Schaar et al., 2008). Following maturation in the Golgi apparatus, virions are released via the host secretory pathway. Kudelko et al. (2012) demonstrated that PrM interacts with the ADP-ribosylation factors 4 and 5 (Arf4 and Arf5), and that depletion of the factors prevents dengue secretion into the extracellular space. Once a virion is released into the extracellular space, the neutral pH of the environment allows dissociation of the Pr peptide (Pierson and Diamond, 2013).

4.1.5 Dengue virus and the unfolded protein response

As detailed earlier in a discussion of the effects of coronavirus infection on cellular gene expression, viral replication can lead to significant changes in host proteostasis that can induce the integrated stress response (ISR) and the unfolded protein response (UPR). Dengue virus replication, translation, protein processing and virion assembly make wide use of the endoplasmic reticulum and knockdown or inhibition of the ER membrane translocon Sec61 significantly suppresses DENV-2 infectivity in human dendritic cells (Heaton et al., 2016).

Pena and Harris (2011) have shown that DENV-2 infection modulates the three arms of the UPR in a time-dependent manner. Dengue infection triggers and then suppresses PERK-mediated eIF2α phosphorylation at early time points. Multiple studies reported that PKR did not play a role in the ISR induction and anti-viral response; PKR is thought to be unable to access the dsRNA intermediates produced during viral replication as they are compartmentalised in DENV induced ER vesicles (Uchil and Satchidanandam, 2003; Welsch et al., 2009). At mid-time time points (24-36 h p.i.) Ire1-Xbp1 pathways are activated, including induction of some downstream targets - GRP78, CHOP and GADD34 but not GRP94 or PDI. Additionally, induction of CHOP did not induce apoptotic markers. Atf6 cleavage and activation only occurred at late time points (48 h p.i.), though mutation studies showed that its role in the dengue-induced UPR was minimal.

The authors of this study propose that early viral protein synthesis triggers the PERK arm of the UPR and transient short-term eIF2α phosphorylation, which is then suppressed by

an unidentified viral protein; the net benefit to the virus being a 'reset' in cellular transcription and translation that allows the virus to utilise more cellular resources for its benefit. The intermediate stage activation of Ire1-Xbp1 is induced by accumulation of viral proteins, and upregulates the chaperone GRP78 which allows transport of the dengue polyprotein into the ER lumen and upregulates GADD34 and GRP78 through CHOP. GADD34 induction leads to dephosphorylation of eIF2α and GRP78 activity inhibits the pro-apoptotic effects of the UPR. In summary, careful manipulation of the UPR is thought to benefit virus replication by freeing resources needed for production of viral components and by increasing expression of various cellular factors that facilitate cross-membrane transport.

Pena and Harris (2012) qualified their first study with laser scanning confocal and differential interference contrast microscopy to show that DENV induces ER rearrangements and expansion at early time points (within 12 h p.i.). Using actinomycin D (an inhibitor of cellular transcription) and cycloheximide, they showed that the structural changes are dependent on production of new viral but not cellular proteins. The rearrangements occurred even in cells where components of the UPR (Xbp1 and Atf6) and SREBP-2 pathways were knocked out.

4.2 Ribosome Profiling of DENV-infected cells

To study the kinetics of virus RNA and protein synthesis, Huh7 cells were infected with Dengue virus type 2 strain New Guinea C (GenBank Accession: AF038403.1). To compare viral replication in a different host, *Aedes albopictus* (Asian tiger mosquito) C6-36 cells were also infected with the same virus. Human cell lysates were prepared at 4, 10, 24 and 30 h p.i. (as well as mock controls at 4 and 30 hours), while mosquito cell lysates were prepared at 24 h p.i. For each condition, RiboSeq (harringtonine), RiboSeq (cycloheximide) and RNASeq libraries were prepared as described previously in Chapter 2. For the *Aedes* cycloheximide library, two replicates were prepared, one in which RPFs were generated with the usual RNase I treatment and one using micrococcal nuclease. Infections were carried out by Dr Andrew Davidson at the University of Bristol. The use of an alternative nuclease was prompted by the work of Dunn and colleagues (2013) who used ribosome profiling to study translation in *Drosophila melanogaster* cells. They found that, in contrast to yeast and mammalian cell lines, *Drosophila* ribosomes are highly sensitive to RNase I, likely due to differences in rRNA folding. However, they were able to use micrococcal nuclease (MNase) as an alternative over a wide range of concentrations. Although MNase does not destroy *Drosophila* ribosomes, it

exhibits a 3′ A/U cutting bias, yielding positional uncertainty for the ribosomal P-site. RNase I shows little cutting bias and yields superior spatial resolution along mRNAs and superior phasing in coding sequences. We chose to investigate the effectiveness of each of these nucleases in our *Aedes* libraries.

The RiboSeq (CHX), RiboSeq (HAR) and RNAseq (CHX only) libraries were deep sequenced using an Illumina HiSeq 2000 platform (Beijing Genomics Institute). Reads were mapped as before, with the exception that reads were mapped directly to the *Aedes* transcriptome, as the mosquito genome and transcriptome were only recently sequenced and assembly of the sequencing contigs and annotation of the genome is still in progress (Chen et al., 2015a).

4.2.1 Data Quality and Library Composition

After mapping the reads to various host and viral read annotations, a significant of unmapped reads remained. These reads were assembled using Trinity and the largest analysed using Blastn. A significant number of the reads aligned to Mycoplasma rRNA. A Bowtie database was generated consisting of the genomes of 108 different Mycoplasma from RefSeq and the alignment procedures repeated. This confirmed that a large portion of sequencing reads in the Huh-7 libraries (between 3.3% and 37%) were derived from mycoplasma contamination (

Figure 4.5, Table 4.2). The low levels of mycoplasma reads present in the C6-36 cell lines were most likely due to generic, non-specific reads that align to the corresponding Bowtie index. A number of unmapped reads remained in our *Aedes* libraries, though these may be due to sequences missing from the current *Aedes* assembly (Figure 4.6).

4.2.2 Read Length and Framing Distributions

Consistent with previously described analyses, RNASeq read lengths show a broad distribution from 29 to 35 nt, while the human cell RiboSeq libraries had a more discrete distribution centered on 30 nt (Figure 4.7).

In the *Aedes* RiboSeq libraries, the RNase I-treated libraries had a similar distribution to the human RiboSeq libraries (

Figure 4.8). However, the MNase-treated RiboSeq library showed a much-wider read length distribution, from 27-35 nt, consistent with previous descriptions of the effect of MNase treatment on ribosome profiling data (Dunn et al., 2013).

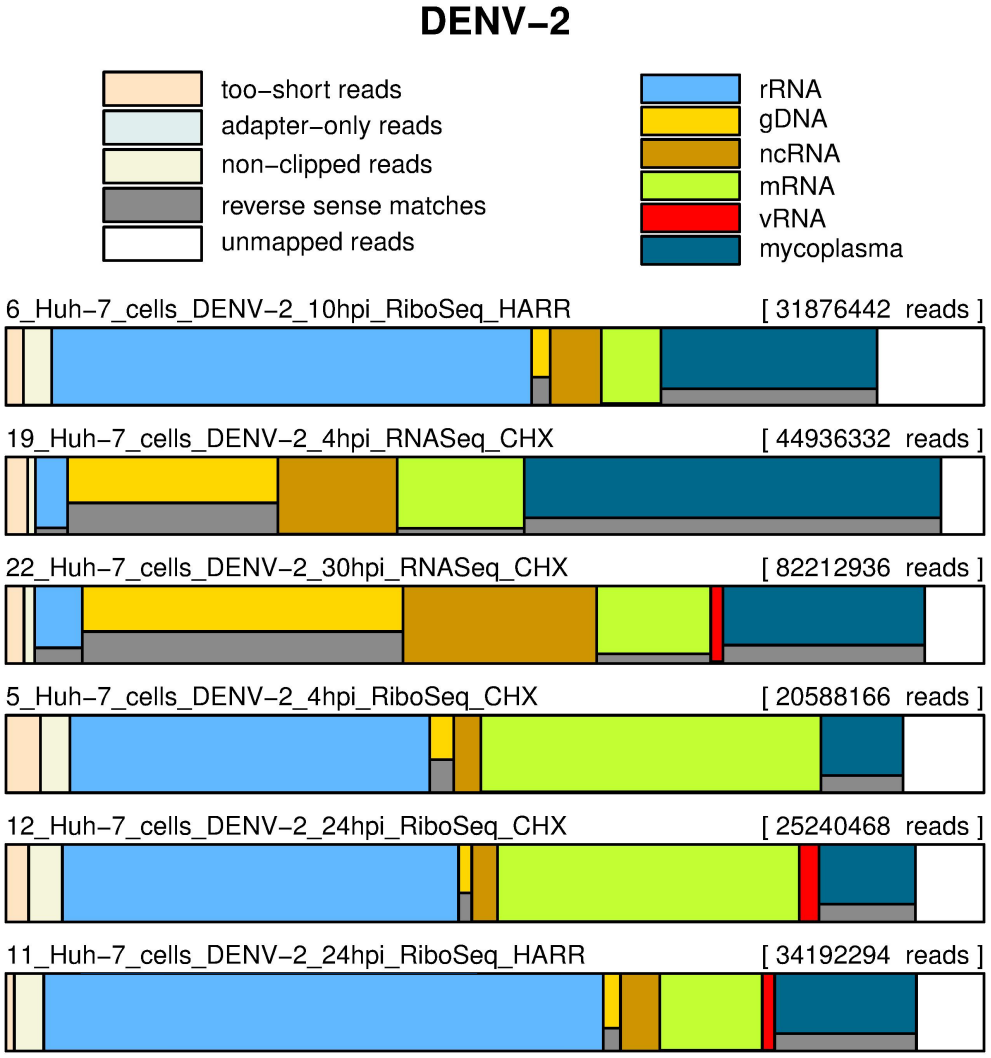


Figure 4.5 Library Composition of select DENV-infected human cell libraries. Selection of DENV-infected and mock libraries from Huh-7 lysates. Some mycoplasma contamination occurred in all libraries, as seen by the number of reads mapping to a Bowtie database assembled from a collection of 108 mycoplasma genomic sequences downloaded from RefSeq.

Library	Total Reads x 1000	Trimmed x 1000	rRNA x 1000	(-) vRNA x 1000	(+) vRNA x 1000	mRNA x 1000	% mRNA	ncRNA x 1000	gDNA x 1000	mycoplasma x 1000
Mock RiboSeq CHX Mnase	35,955	35,346	15,156	0	0	8,099	18.38%	0	4,042	11
Mock RNASeq CHX	39,097	38,190	16,541	0	0	2,003	4.87%	0	12,747	24
Inf RiboSeq CHX MNase	32,563	32,243	27,675	0	111	784	2.35%	0	999	4
Inf RiboSeq HARR RNaseI	31,361	31,057	27,483	0	14	30	0.10%	0	1,184	5
Inf RiboSeq CHX RNaseI	93,624	92,092	79,751	1	96	852	0.90%	0	4,175	21
Inf RNASeq CHX	30,122	29,678	6,223	66	451	2,158	6.69%	0	13,163	6
4hpi RiboSeq HARR	20,032	17,279	7,992	0.042	1	1,903	8.67%	646	357	3,303
4hpi RiboSeq CHX	20,588	19,243	7,575	0.030	1	7,155	25.79%	574	505	1,726
4hpi RNASeq CHX	44,936	43,584	1,475	0.179	1	5,842	11.50%	5,479	9,662	19,143
10hpi RiboSeq HARR	31,876	30,386	15,642	0.359	4	1,938	5.73%	1,664	609	7,041
10hpi RiboSeq CHX	19,870	18,256	6,677	0.099	5	6,279	24.01%	762	429	2,974
10hpi RNASeq CHX	64,616	63,086	2,967	1	3	7,189	10.01%	10,867	15,543	23,449
24hpi RiboSeq HARR	34,192	32,870	19,553	5	429	3,575	9.47%	1,378	609	4,955
24hpi RiboSeq CHX	25,240	23,784	10,224	2	508	7,784	23.57%	663	341	2,491
24hpi RNASeq CHX	53,822	52,605	2,109	16	306	7,258	11.88%	9,386	15,140	15,020
30hpi RiboSeq HARR	13,660	11,716	6,483	4	104	894	6.14%	475	326	2,145
30hpi RiboSeq CHX	17,417	14,752	7,310	6	274	2,327	11.79%	529	340	2,595
30hpi RNASeq CHX	82,213	79,792	4,008	28	1,006	9,586	10.44%	16,279	26,946	16,963
4hr Mock RiboSeq CHX	33,437	33,053	25,444	0.002	0.060	3,737	10.05%	639	273	1,105

4hr Mock RNASeq CHX	27,350	26,808	740	0.004	0.098	2,918	9.64%	4,842	6,703	10,230
30hr Mock RiboSeq CHX	24,337	21,361	8,670	0.002	0.113	7,912	24.53%	951	420	2,219
30hr Mock RNASeq CHX	65,997	62,114	2,330	0.070	2	7,603	10.33%	14,860	19,973	14,045

Table 4.2 Number of reads in each RiboSeq and RNASeq library. Raw read counts for each library and number of alignments stratified by database type. Pink labels indicate C6-36 (*Aedes albopictus*) libraries while blue labels indicate Huh-7 libraries. The human cell line libraries showed significant levels of mycoplasma contamination. The low levels of mycoplasma reads present in the C6-36 cell lines are most likely due to generic, non-specific reads aligning to the corresponding Bowtie index. The lack of a publicly available ncRNA annotation for *Aedes albopictus* prevented the mapping and quantification of reads originating from this source. In the DENV infected libraries, the ratio of (-) to (+) sense virus reads was highest at 4 h p.i. and then dropped off during successive time points.

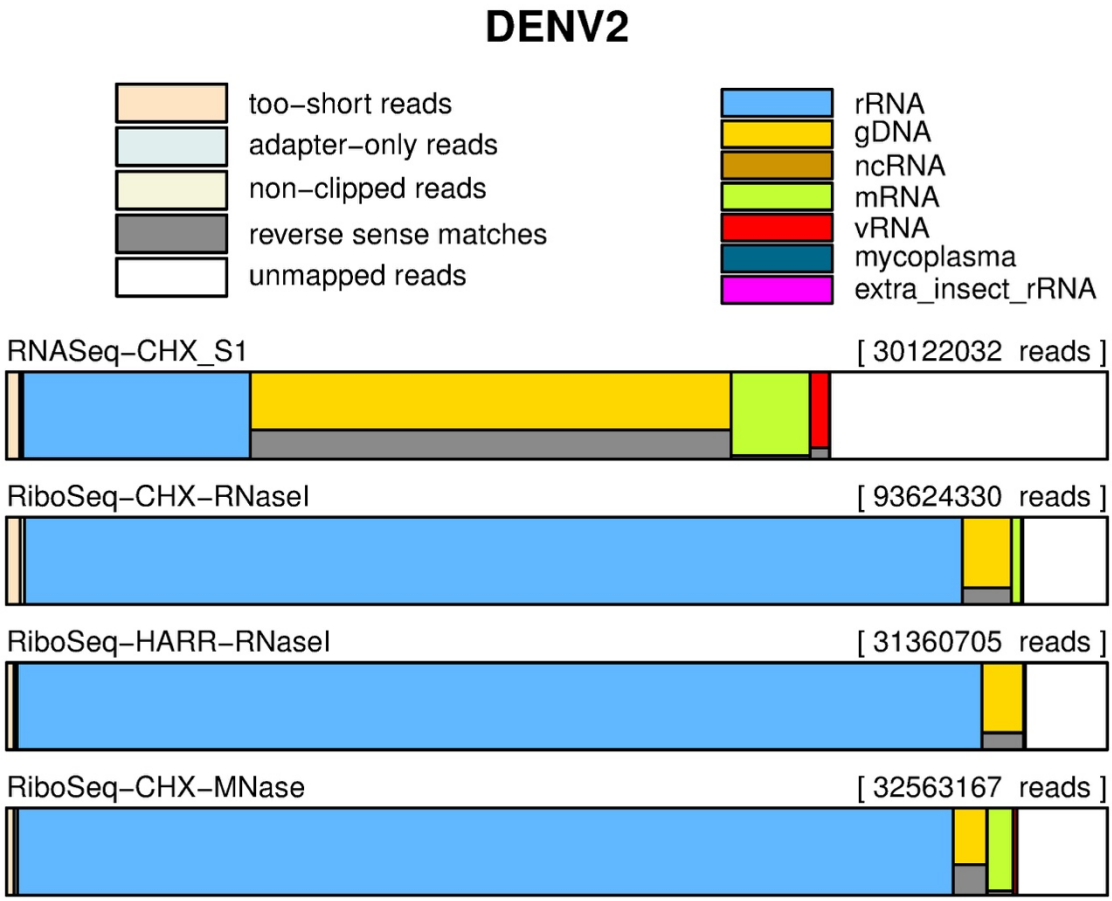


Figure 4.6 Library composition of selected DENV-infected *Aedes* cell libraries. MNase digestion of *Aedes albopictus* lysates yields higher level of mRNA reads relative to RNase I treatment. The 'extra insect rRNA' category includes rRNA from several other arthropod species for which rRNA sequences are available on RefSeq. This extra alignment step was done to check whether any of the unmapped reads at the end of pipeline could originate from insect rRNA that is not currently annotated in the present *Aedes albopictus* assembly.

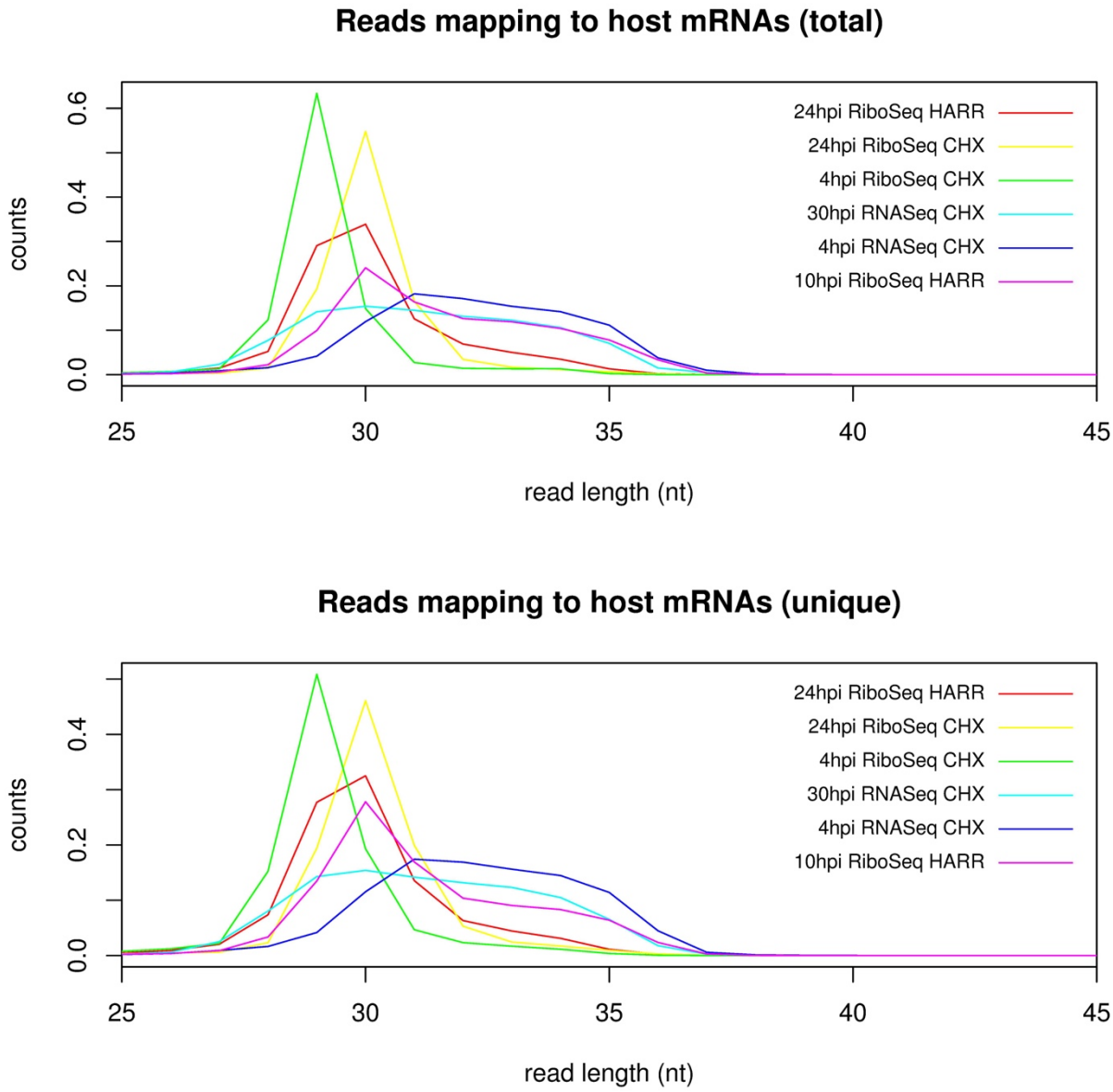


Figure 4.7 Read length distribution for human cell libraries. Relative proportions of reads mapping to host mRNAs in Huh-7 libraries. RiboSeq libraries reveal a discrete read length distribution centred on 30 nt while the RNASeq libraries show a much broader read length distribution.

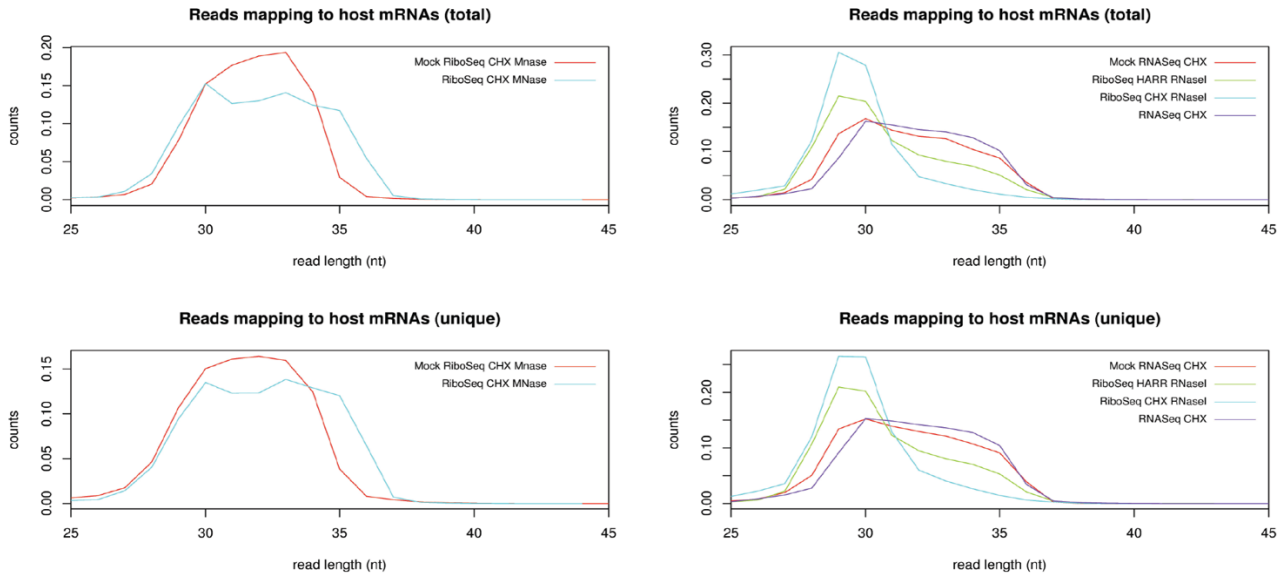


Figure 4.8 Read length distribution in *Aedes* libraries. Relative proportions of reads with each length mapping to host mRNAs in various *Aedes* libraries. MNase treatment results in a wider distribution of RiboSeq read lengths relative to RNase I treatment.

Similarly, the phasing distributions in the RNase I-treated *Aedes* replicates show a significant trinucleotide periodicity, while RPFs from the MNase-treated libraries are much more heterogeneous in their phasing (Figure 4.9). The MNase-treated samples yielded a higher proportion of cellular mRNA reads relative to the RNase-treated samples, though given the high levels of rRNA contamination present in these libraries (despite DSN-treatment), we cannot conclusively say whether one nuclease will consistently yield higher levels of RPFs (Figure 4.6, Table 4.2). Both nucleases yielded comparable levels of vRNA RPFs.

4.2.3 Kinetics of DENV replication

Consistent with previous studies, the RNASeq data from DENV-infected Huh7 cells reveals that at early time points, ~5-10% of the viral RNA is the negative-sense transcripts, and the ratio of -/+ RNA diminishes at later time points as positive-strand amplifications takes place (Figure 4.10). Low levels of translation were observed at the early time-point that increased by several orders of magnitude at later time points as the virus entered into the 'translation-focused' stage of its replicative cycle (Bäck and Lundkvist, 2013).

Sliding window (averaged read density over a range of positions) analyses of the DENV RiboSeq and RNASeq alignments revealed two features (Figure 4.11). Firstly, there was roughly uniform RPF coverage of the entire polyprotein, thus there was no evidence to support significant ribosomal drop-off as a means of downregulating replicase protein expression. Secondly, we saw substantial RNASeq density in the last ~0.4 kb of the 3′ UTR, consistent with production of the sfRNA. However, we found no evidence for translational activity in the 3′ UTR or on the sfRNA.

4.3 Pause sites in the DENV translatome

Alignment of RiboSeq and RNASeq reads to the DENV RNA reveals several features (Figure 4.12). Firstly, we do not see any evidence for the use of the upstream AUGs in the 5′ UTR that could lead to an N-terminally extended polyprotein (Figure 4.13). Several sites of obvious ribosome pausing were evident, the two most prominent residing in the NS2A and NS5 encoding portions of the ORF. In the initiation profiling library (HAR), we note a sparsity of reads mapping immediately downstream of the canonical AUG up until the terminal portion of the NS1-encoding segment of the ORF reflecting the harringtonine inhibition of initiation but

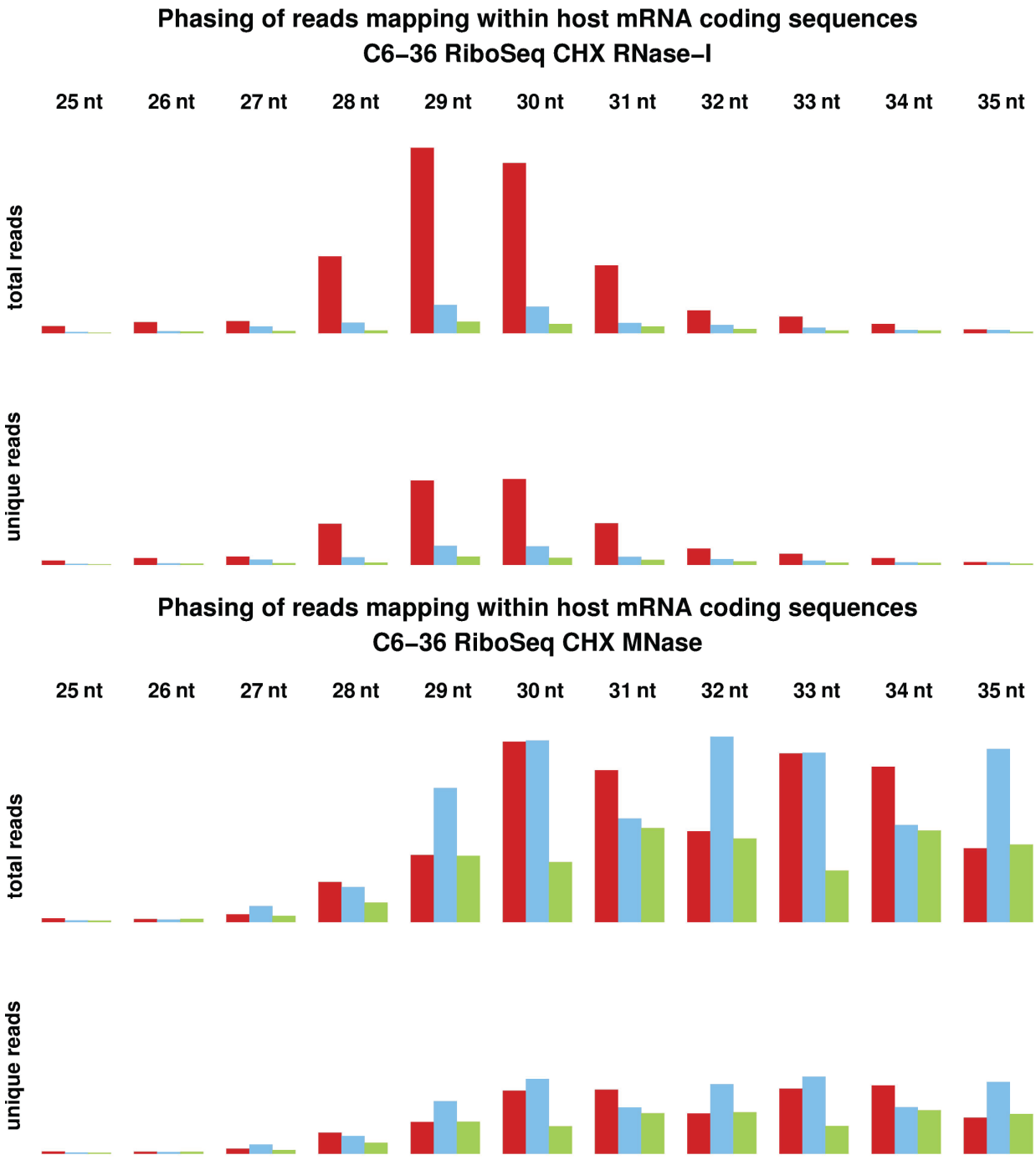


Figure 4.9 RPF phasing distributions in *Aedes* RNase I and MNase treated libraries. Relative levels of reads mapping to each phase of a given CDS, stratified by read length, for two DENV-infected *Aedes* libraries. The top panel shows the effect of RNase treatment, which gives a much more discrete three-nucleotide periodicity, while the lower panel shows the phasing from MNase-treated cells, which gives a much broader distribution.

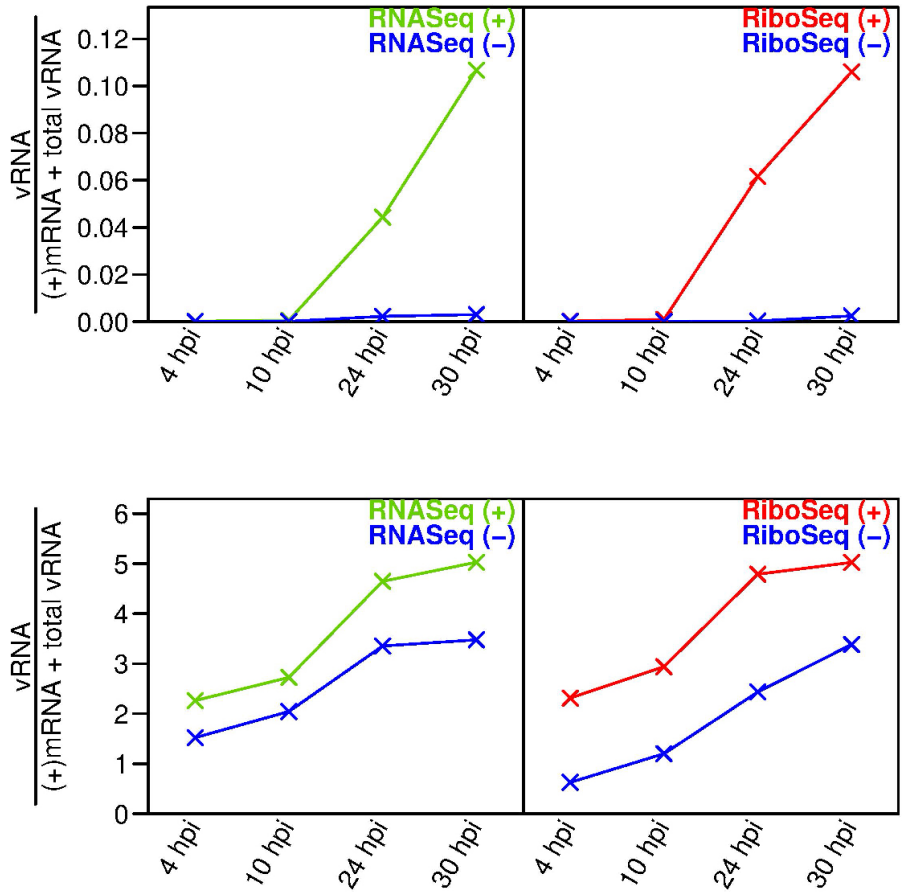


Figure 4.10 Time course of viral RNA synthesis. The top panel is labelled with a linear scale while the bottom panel is labelled with a logarithmic scale for comparison.

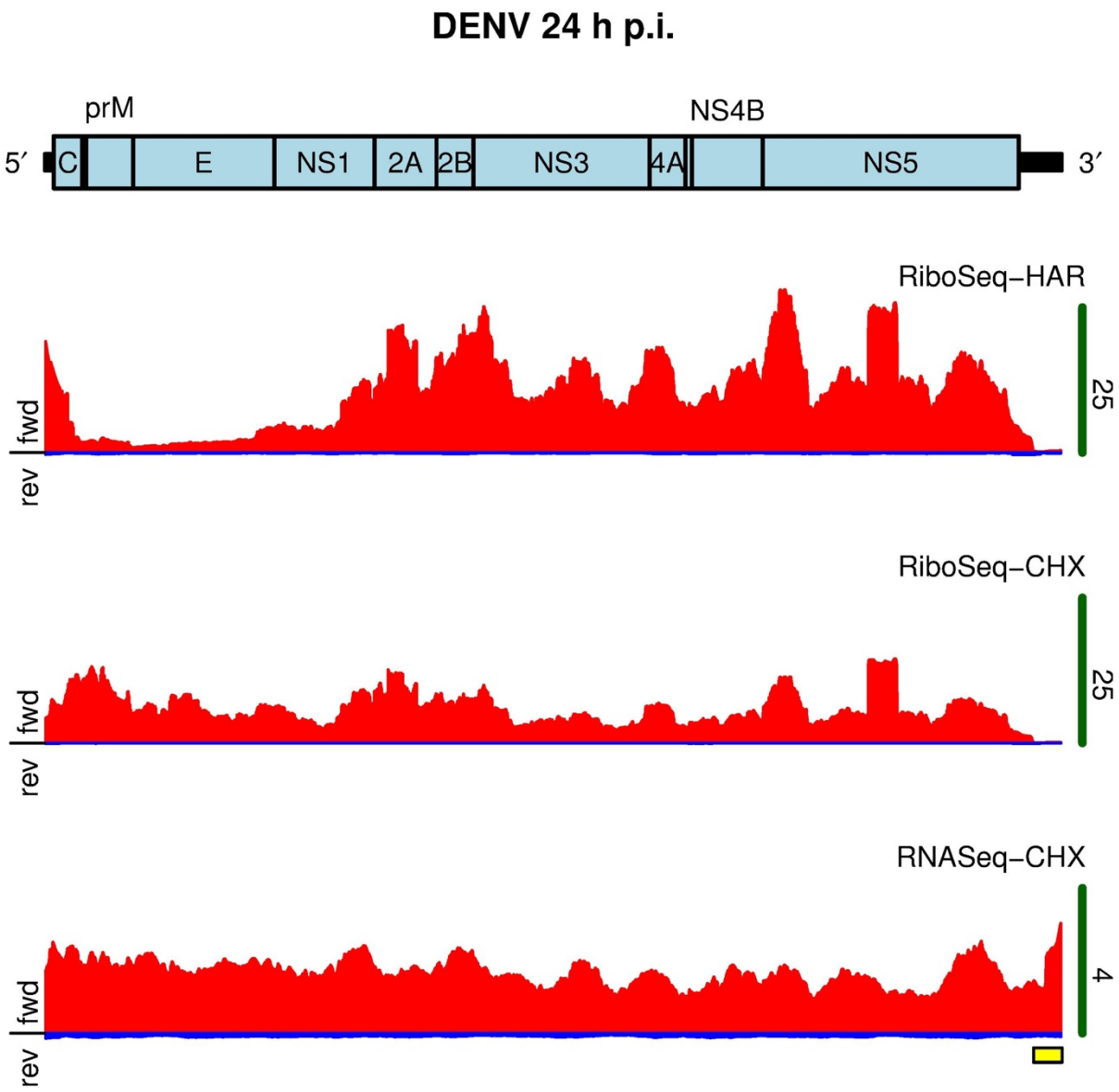


Figure 4.11 300-nt sliding window plot of DENV RNA translation in human cells at 24 h p.i. Translation efficiency of the DENV ORF appears to be fairly uniform. The peak in RNASeq coverage over the last ~0.4 kb of the 3′ UTR reflects production of the sfRNA (sfRNA region indicated by yellow box).

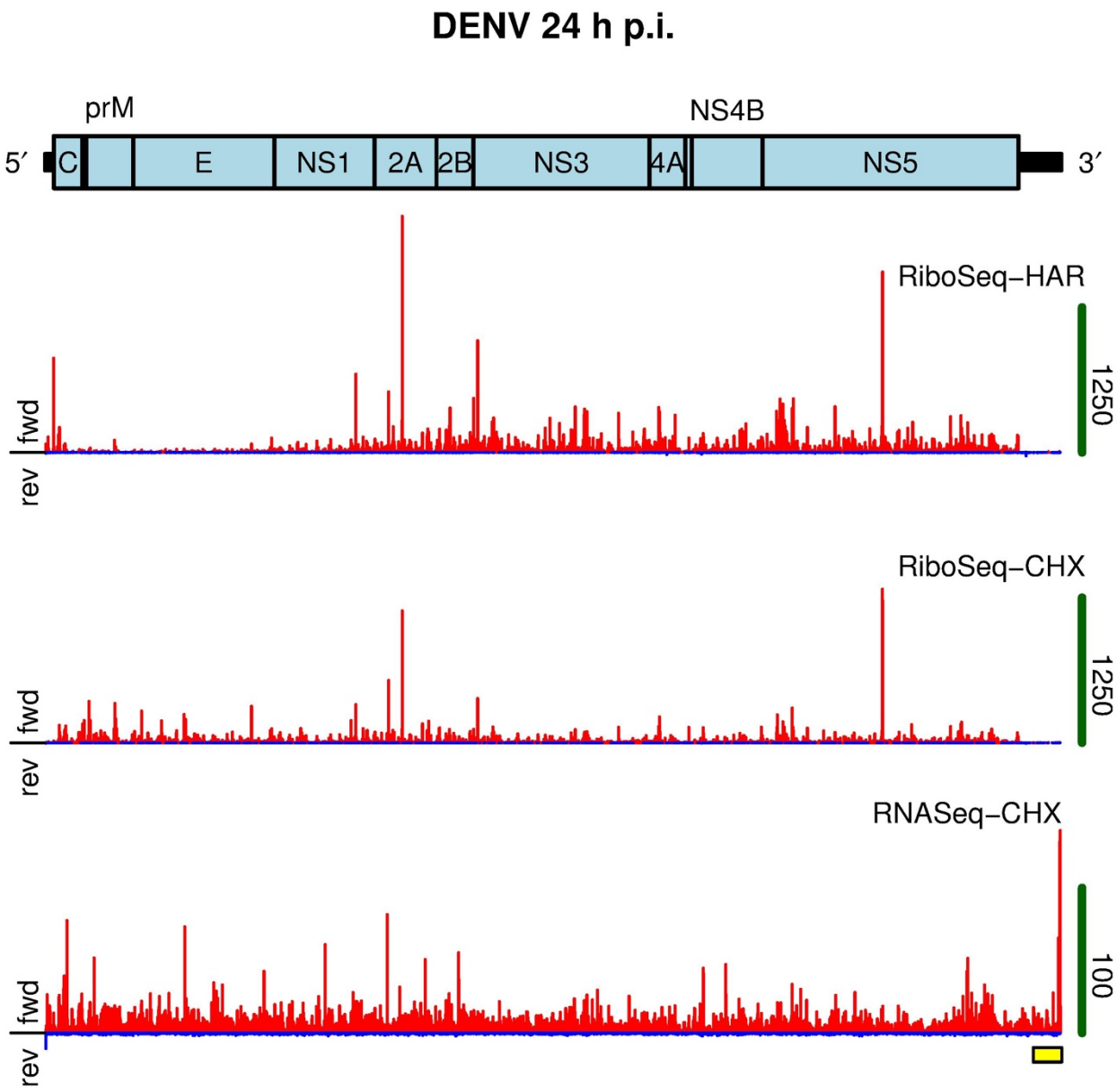


Figure 4.12 Genome map of DENV translation in human cells at 24 h p.i. RiboSeq and RNASeq reads aligning to the DENV RNA in the Huh-7 derived libraries at 24 h p.i. Major RiboSeq pause sites occur in the NS2A and NS5 portions of the ORF, which are shown in greater detail in Figure 4.16 and Figure 4.17. The spike in RNASeq reads in the 3′ UTR may originate from the sfRNA or a partial exonuclease degradation product.

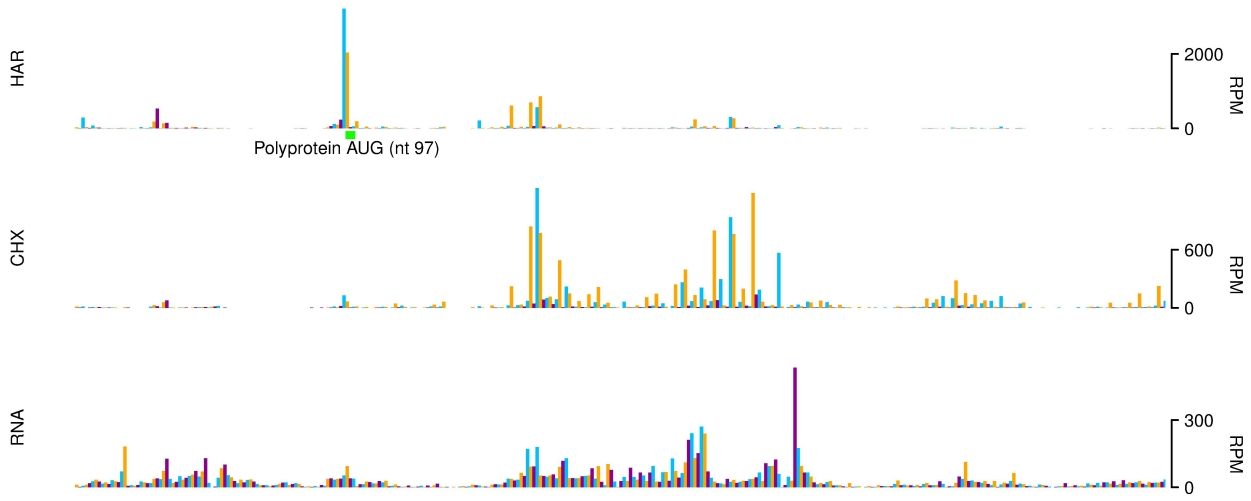


Figure 4.13 Zoom-in plot of DENV ORF start codon. Reads aligning to DENV RNA positions 1-350 in the human 24 h p.i. replicate. Reads whose 5′ ends map to the first, second or third phase relative to position 1 in the DENV genome are indicated in blue, yellow or purple, respectively.

not of continued ribosome elongation. Comparison of the human library-derived read alignments with those from the *Aedes* libraries revealed some differences in pause sites, suggesting possible host species-specific differences in DENV RNA conformation or RBP interactions. We observed a significant peak in the RiboSeq data in the 3′ UTR that only occurs in the *Aedes* dataset (Figure 4.12, Figure 4.14, Figure 4.15).

The first RPF peak in the NS2A ORF occurs at position 3627 (using a +12 nt offset for the 5′ end) (Figure 4.16) and the most common read mapping to this position is AACAUGUCCUUUAGAGACCUGGGAAGAGUG (P-site codon underlined). The corresponding upstream peptide is VAVSFVTLITGNMSFR. The second, stronger RPF peak in the NS2A ORF occurs at position 3771, with the most common read mapping to this position being AAGUUGACCUCCAAGGAAUUGAUGAUGACU (P-site codon underlined), with corresponding upstream peptide VRPTFAAGLLLRKLTSK. The RPF peak in the NS5 ORF occurs at position 8835 (Figure 4.17), with the most common read mapping to this position being UGGAAGUCGGCACGUGAGGCUGUUGAAGA and the corresponding upstream peptide SNAALGAIFTDENKWKSAR. The last two reads have a glutamic acid encoding codon present in the A-site, consistent with a previous profiling report that noted an enrichment of aspartate or glutamate codons in the A site at strong pause sites (Ingolia et al., 2011). RNA folding analysis using pKnots RG did not identify any strong secondary structures downstream of these pause sites (Janssen and Giegerich, 2014). However, analyses of the NS2A peptide sequences using the hidden Markov model prediction suite for transmembrane helices TMHMM (Krogh et al., 2001) revealed that each of the two pause sites in NS2A occurs at the corresponding C-terminal portions of predicted transmembrane helices (though each nascent helix would still be within the ribosome exit tunnel). The TMHMM analysis of the NS2A also suggested the presence of a third transmembrane segment between amino acids 133 and 155, though no major pause site was observed at this location.

The 3′ UTR RiboSeq peak in the *Aedes* libraries occurs at position 10500 (Figure 4.18), and the most common read aligning to this position was GUAGUGGACUAGCGGUUAGAGGAGACCCCUCCC. The read length distribution for this position were noted to be 24.8% 34 nt reads, 68.5% are 33 nt, 5.7% 32 nt and the remaining ~0.01% were 26-31 nt in length, making this read unusually long for the RNase I libraries. This portion of the 3′ UTR is thought to be processed into part of the sfRNA. Given the unusual read distribution for this peak, it is likely that these reads are derived from an RBP interacting with the sfRNA (and the RNP co-sedimenting with ribosomes) or due to the effect of one of the numerous secondary structures present in the 3′ UTR.

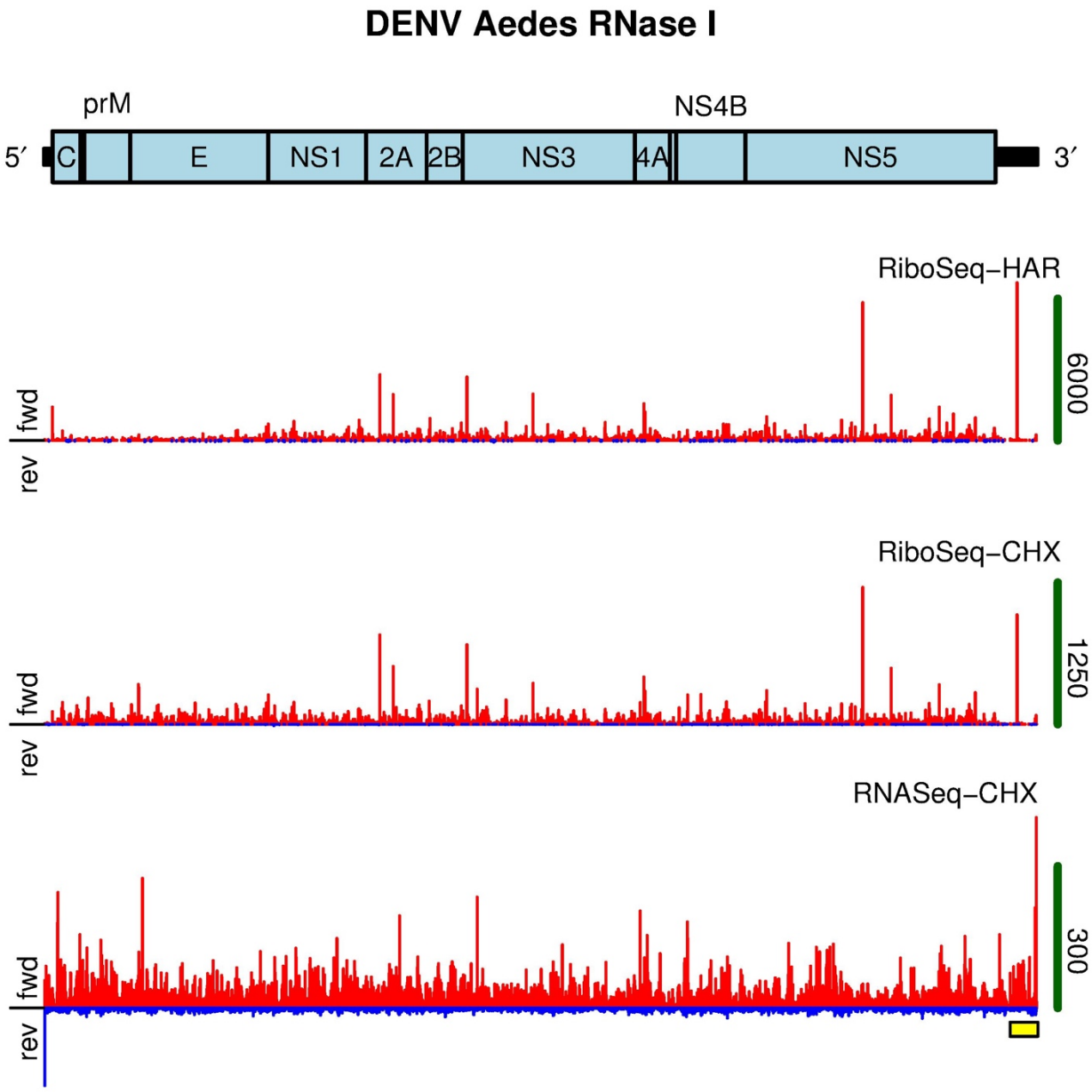


Figure 4.14 Genome map of DENV translation in C6-36 cells with RNase I treatment for RiboSeq libraries. RiboSeq and RNASeq reads aligning to the DENV RNA in the C6-36 cell line derived libraries at 24 h p.i. The RiboSeq cycloheximide library in this figure was treated with RNase I.

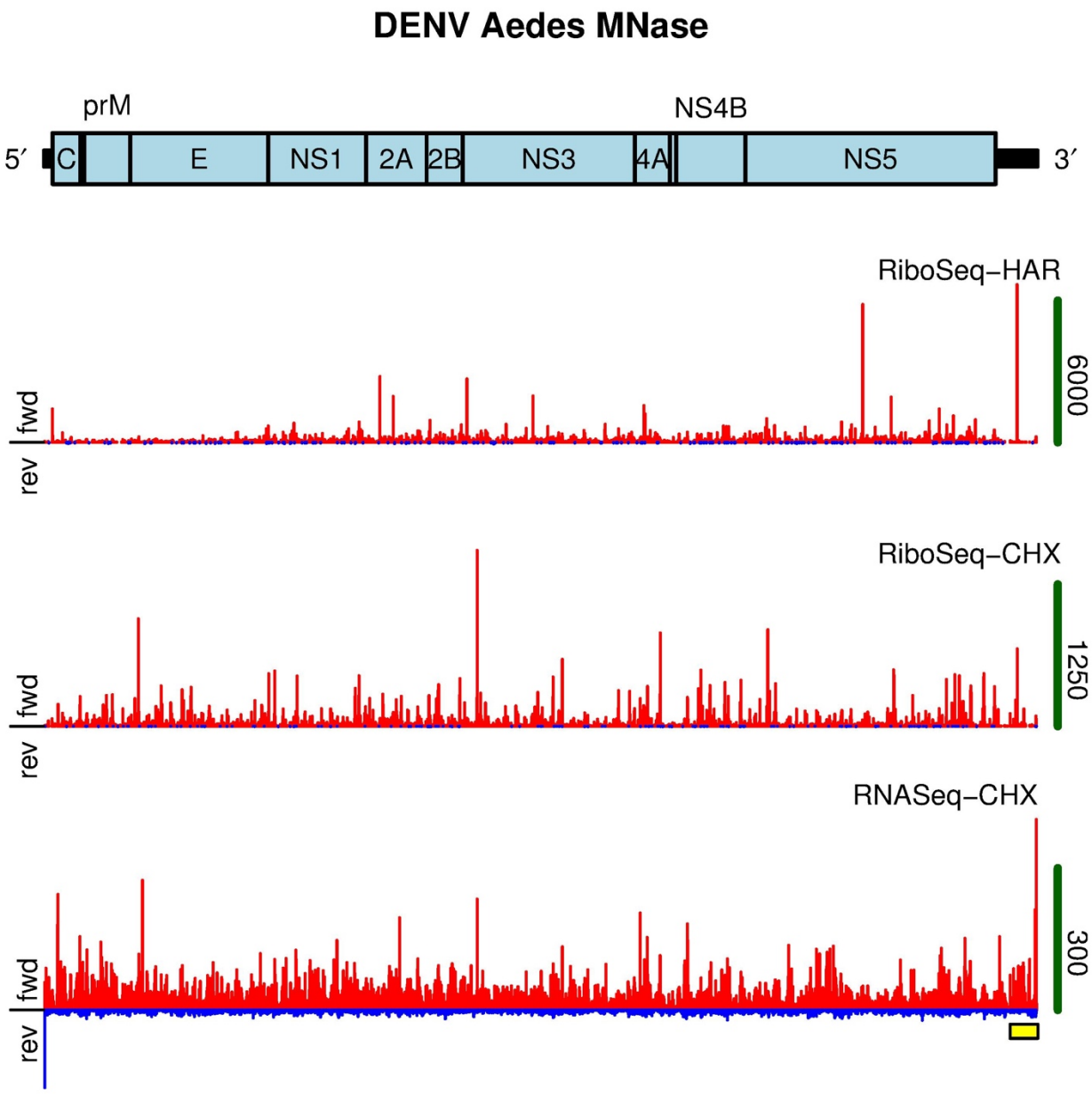


Figure 4.15 Genome map of DENV translation in C6-36 cells with RNase I treatment for RiboSeq libraries. RiboSeq and RNASeq reads aligning to the DENV RNA in the C6-36 cell line derived libraries at 24 h p.i. The RiboSeq cycloheximide library in this figure was treated with MNase.

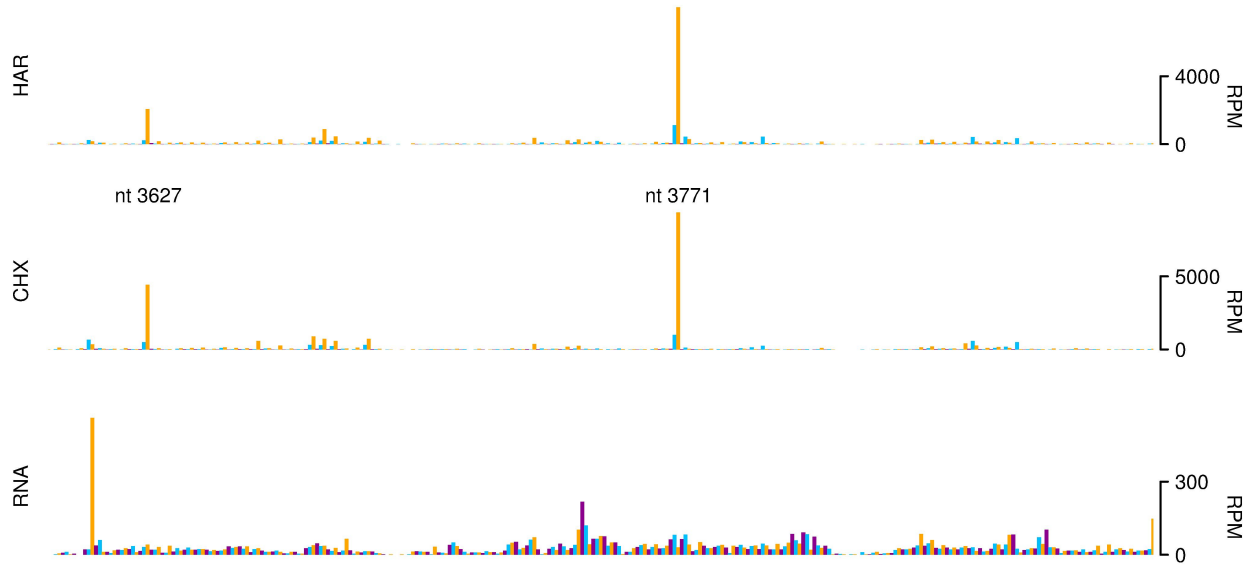


Figure 4.16 Zoom-in plot of the NS2A-encoding portion of the DENV ORF. Reads aligning to positions 3600-3900 in the DENV RNA in the 24 h p.i. Huh-7 library. The peak at position 3627 occurs at NS2A codon 38, at the C-terminal end of a predicted transmembrane helix (amino acids 21-40); the peak at position 3771 appears at codon 82, the end of a second C-terminal transmembrane helix from amino acids 60-82, as predicted by TMHMM (Krogh et al., 2001).

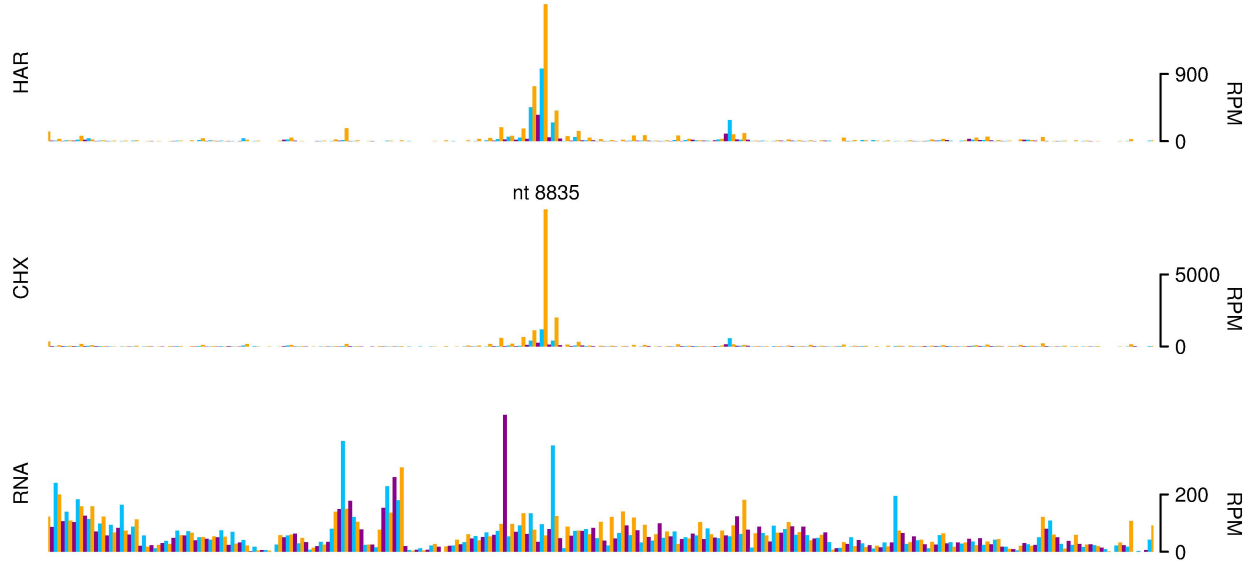


Figure 4.17 Zoom-in plot of NS5-encoding portion of the DENV ORF. Reads aligning to positions 8700-9000 in the DENV RNA in the 24 h p.i. Huh-7 library. RNA folding analysis using pKnots RG did not identify any strong secondary structures downstream of the RPF peak at position 8835, nor was there any obvious potential for peptide mediated stalling.

Figure 4.18 Zoom-in plot of the DENV 3′ UTR. Reads aligning to positions 10240-10724 in the DENV RNA in the 24 h p.i. *Aedes* RNase-treated library (top panel) and in the 24 h p.i. human library (bottom panel). Position 10500 is within the ~0.4 kb portion of 3′ UTR thought to be preserved during Xrn1 degradation to form the sfRNA.

4.4 Discussion

To our knowledge, this study represents the first high-throughput analysis of flavivirus translation. The results corroborate previous work showing that cHP induces the vast majority of ribosomes to initiate at the 'canonical' AUG of the DENV ORF (Clyde and Harris, 2006). We confirmed expression of the sfRNA in both human and mosquito cells and show that various pause sites are host species-specific, though biological repeats are needed to are needed to confirm that these observations are not the result of a technical artefact or random noise. Several studies have shown that certain portions of the DENV 3′ UTR have species-specific signals for viral replication (Alvarez et al., 2005; Tajima et al., 2007; Blaney et al., 2008; Villordo and Gamarnik, 2013). Alternative methodologies such as CLIP-Seq or SHAPE will need to be employed to investigate the specific causes of RPF accumulation around these various sites. Though we observed major RPF peaks in the vicinity of two predicted transmembrane helices in the NS2A region, we did not observe a major peak at the site of a third predicted NS2A transmembrane segment, nor at the transmembrane segments of several of DENV proteins.

Certain viruses, such as encephalomyocarditis virus, have been shown to downregulate expression of their polymerase proteins via translational readthrough or frameshifting (Ahlquist, 2006). We hypothesized that NS5 and perhaps other C-terminal portions of the DENV polyprotein might be translated with reduced efficiency, but actually we did not observe this in the profiling datasets. It is possible that through controlled compartmentalization of the various replicase proteins, such as by phosphorylation-dependent nuclear import of NS5, DENV can avoid inducing cytotoxic effects while producing its structural and non-structural proteins at similar levels.

The current study was unfortunately frustrated by mycoplasma contamination in the human cell cultures, though we were able to examine the feasibility of using MNase and RNase I in arthropod cell lines. Similarly to Dunn et al. (2013), we found that RNase I treatment produces superior read phasing but a smaller proportion of cellular mRNA RPFs than the MNase treatment. However, despite DSN treatment, both MNase and RNase I cleaved samples had high levels of rRNA contamination, so we cannot definitively conclude whether MNase treatment would consistently yield a higher proportion of cellular RPFs. The relative levels of DENV vRNA RPFs in each replicate were roughly the same. A control experiment using a more extensive rRNA depletion step (such as with a RiboZero kit) could help discern the relative utility of each nuclease. In future arthropod ribosome profiling analyses, it may be optimal to use MNase derived datasets for differential expression analyses or detection of low-

abundance RNAs, and RNase I treatment for analyses of short ORF translation and codon utilization.

Given that our lysates were prepared without compartment fractionation and that some DENV RNA is in a translation competent form while other molecules are packaged or used for RNA replication, we were unable to calculate relative DENV translation efficiency (i.e. DENV TE relative to host mRNA TE). Though DENV is thought to inhibit the formation of stress granules, it makes use of virally induced vesicle compartments for replication, which may affect RNA experimental isolation. Additional work will be needed to control for the effect of these compartments on RNASeq and RiboSeq analyses of DENV-infected cells.

Future investigations (in mycoplasma-free cells) will repeat these experiments to confirm the occurrence of various pause sites as well as to allow for differential expression analyses of host translation and transcript abundance (particularly the role of various components of the UPR and ISR pathways). These studies will also help quantify the relative host and viral utilization of the cellular translation apparatus, and potentially identify virus-specific mechanisms that can be exploited for therapeutic intervention.

Chapter Five

Multi-mapping reads in ribosome profiling analyses

5 CHAPTER FIVE: MULTI-MAPPING READS IN RIBOSOME PROFILING ANALYSES

5.1 Introduction

Current NGS platforms produce reads that are typically much shorter than the native nucleic acids molecules from which they are derived (Goodwin et al., 2016). While the development of novel sequencing methods (such as PacBio SMRT or Oxford Nanopore) may allow for the sequencing of full length, unfragmented mRNA molecules, these technologies are still undergoing optimization and are not in wide-spread use (Buermans and den Dunnen, 2014; Feng et al., 2015). For example, SMRT suffers from a high error rate and relatively low throughput which hinders quantification of transcript expression, particularly for low abundance transcripts (Au et al., 2012; Conesa et al., 2016). Certain applications, such as the study of short non-coding RNA or ribosome profiling, will always require the use of short sequencing reads.

Shortly after the development of RNASeq and the tools to handle mapping of millions of short reads generated during typical Illumina sequencing experiments, it became apparent that certain reads could not be unambiguously assigned to one unique genetic locus. The use of shorter read fragments or allowing a small number of mismatches in alignments invariably increases the proportion of reads that cannot be uniquely mapped to one particular site in a cellular genome. Many studies have handled this issue simply by discarding multi-mapping reads, leading to a potential loss of biological data. Other studies have chosen to simply keep each potential alignment for a multi-mapping read in their computational analyses, leading to artificially inflated read counts for certain genetic elements.

This chapter will summarize previous work on non-unique sequence elements (primarily in the context of RNASeq), introduce several analyses to explore the implication of repetitive sequences particularly in coding sequences and the effects on ribosome profiling experiments, and lastly suggest potential ways of mitigating the issue of repetitive sequences during RiboSeq projects.

5.1.1 Previous work on multi-mapping reads and RNASeq

One of the first papers to describe the use of NGS in the study of a mammalian transcriptome noted how a significant portion of reads could not be uniquely assigned, even when disregarding the issue of isoform quantification (Mortazavi et al., 2008). The authors of this study performed RNASeq on poly(A)-selected RNA from C57BL mouse brain, liver and skeletal muscle tissues,

generating 10-30 million 25-bp reads mapping to unique sites per sample. After simulating the mouse muscle transcriptome, they found that 76% of 25-bp fragments had a unique mapping, 21% had 2-10 potential mappings, and 3% of fragments had 11+ mappings. Many of the multireads in these datasets were attributable to known duplicated genes and segmental duplications. The Ubiquitin B family was shown to be dominated by multireads (97%), as expected for paralogs that are very similar to each other and for internally repeated domains within some genes.

Simulations of 50-nt reads in *Arabidopsis*, *Brachypodium*, maize, potato, rice, soybean and tomato allowed the identification of plant genes recalcitrant to RNAseq analyses by having over- and/or under-estimated expression levels (Hirsch et al., 2015). In maize, over 25% of genes deviated by more than 20% from the expected simulated count values, suggesting the need for cautious interpretation of RNAseq data for certain elements. When two mismatches are allowed with the Bowtie short read aligner, 17% of mouse sequences and 52% of maize sequences are potentially multi-mapping (Li et al., 2009a).

Chung and colleagues (2011) studied the issue of multi-mapping reads in the context of ChIP-seq, and how discarding these reads can contribute to a high level of false negatives. Previous ChIP-seq analyses used only reads that could be mapped uniquely to a reference genome, leading to omission of an estimated 30% of alignable reads in six of these studies. These authors allocated multi-reads as fractional counts using a weighted alignment scheme, and illustrated the approach on human STAT1 and mouse GATA1 ChIP-seq datasets. The new methodology detected novel peaks that were not otherwise identifiable with unique reads, particularly in low complexity segments or in recent gene duplication. The study identified novel target genes of GATA1 and increased the number of predicted binding regions by up to 36%.

A recent comparative analysis of 12 methods for RNASeq read alignment and quantification identified hundreds of genes whose expression is underestimated by one or more methods, and that some of these genes are of clinical relevance (Robert and Watson, 2015). DAZ, a gene whose deletion has been correlated with infertility in South Chinese and Tunisian males, and RPSMY1 (responsible for regulation of sperm motility) were severely undercounted in these simulations. Several cancer/testis (CT) family genes (commonly expressed in tumours), including CT47A (involved in X-linked mental disability), did not have detectable expression levels when only uniquely mapping reads were used. Several GAGE genes (cancer antigens expressed in many tumours) were underestimated by 50%. UTY genes, which have been correlated with increased risk of coronary artery disease, also contain many multi-mapping

reads. The study used various combinations of alignment programs Star, Tophat and SailFish, and read quantification algorithms HTSeq-count or Cufflinks. Each of these approaches lead to underestimation and/or overestimation of read counts for certain genes. The authors of this study propose a 2-stage analysis of RNASeq data in which unique reads are mapped first and then reads that map to multiple genes are assigned uniquely to "multi-map groups" (MMGS). Using a previously published mouse cancer study and edgeR, 672 additional genes groups were identified as differentially expressed relative to when the analysis was done only using uniquely mapping reads.

Besides creating gene groups to classify elements with high level of sequence similarity, several other approaches have been proposed for dealing with multi-mapping RNASeq reads. The Cufflinks isoform quantification algorithm will take reads for which there are *n* equally valid alignments and assign a probability of 1/n for each alignment being valid (Trapnell et al., 2010). The SAM format encodes this probability in the mapping quality field, and is later used by Cufflinks during isoform FPKM calculations. The MMR algorithm infers optimal mapping locations from the coverage density of other mapped reads, making use of the critical fraction of unambiguously aligned reads and iteratively selects the alignments of ambiguously mapping reads in a way that the overall coverage becomes more uniform (Kahles et al., 2015). The algorithm selects between competing mapping possibilities by optimizing for local alignment density.

Unlike other heuristic approaches which rescue multi-mapping reads by allocating fractions of them to genes in proportion to coverage by uniquely mapping reads; Li and colleagues (2009) developed the RSEM algorithm to estimate maximum likelihood (ML) expression levels using an Expectation-Maximization (EM) algorithm. RSEM first preprocesses a set of reference transcript sequences and then aligns a set of RNASeq reads to reference indices. The resulting mappings are used to estimate abundances and credibility intervals and produces probability-weighted alignments in the BAM format. RNASeq simulations showed that RSEM performs favourably in terms of sensitivity compared to Cufflinks and other algorithms (Li et al., 2009a).

5.1.2 Repetitive sequences and ribosome profiling analyses

Multi-mapping reads are a particular problem in the case of ribosome profiling experiments, as these experiments necessitate the use of ~30 nt read fragments due to the

that the ribosome mRNA tunnel protects transcripts from nuclease degradation. While analyses have explored the potential for RNASeq reads to map to different locations in the transcriptome, the extent of this artefact has not been formally documented in the case of translatome analyses, where the vast majority of reads are expected to derive from only the coding portions of transcripts and read lengths are narrowly distributed around 30 nucleotides. Discarding multi-mapping reads can lead to loss of information on the translation of particular sets of genes, while allowing for multiple alignments can skew RPF distributions and increase Poisson noise that leads to aberrant differential expression calls (

Figure 5.1).

Mammalian histone transcripts share high levels of sequence similarity between portions of their ORFs, as is the case for several other paralogous gene families that likely arose through gene duplication events (Figure 5.2; Huminiecki and Wolfe, 2004). Allowing for multiple alignments for the datasets described in Chapter 3 yields significantly inflated read count measurements for multiple murine histone transcripts (Figure 5.3, Figure 5.4). Reads likely derived from other histone ORFs 'stack' at shared sequence sites, leading to localized regions with exaggerated RPF alignment levels.

Several published ribosome profiling studies have failed to account for the potential multi-mapping reads (

Table 5.1). Lee and colleagues (2012) sought to map translation initiation sites (TIS) in mammalian cells using ribosome profiling and a combination of cycloheximide (CHX) and lactimidomycin (LTM) treatment. The study sought to "reduce the background noise of LTM-associated RPFs further by subtracting the normalized density of CHX reads at every nucleotide position from the density of LTM reads at that position. A TIS peak then is called at a position in which the adjusted LTM reads density is well above the background." The reads in this study were mapped allowing up to 100 potential alignments. Retention of multi-mapping reads would likely inflate the RPF density at certain transcript positions in the cycloheximide-treated libraries and potentially reduce the detection of alternate translation initiation sites.

Another profiling study in *C. elegans* mapped reads to a reference index containing one isoform per gene, reporting all alignments with perfect mismatches or a single mismatch (Stadler and Fire, 2013). A study of m6A-dependent regulation of mRNA stability in human cells mapped reads using Tophat without any gaps and allowing for at most two mismatches (Wang et al., 2014). The default Tophat (and Bowtie) parameters allow for up to 200 alignments per read before a read is discarded. One report used ribosome profiling to quantify the

correlation between various NMD inhibition treatments such as *Upf1* knockdown with translation efficiency values for different transcripts (Hurt et al., 2013).

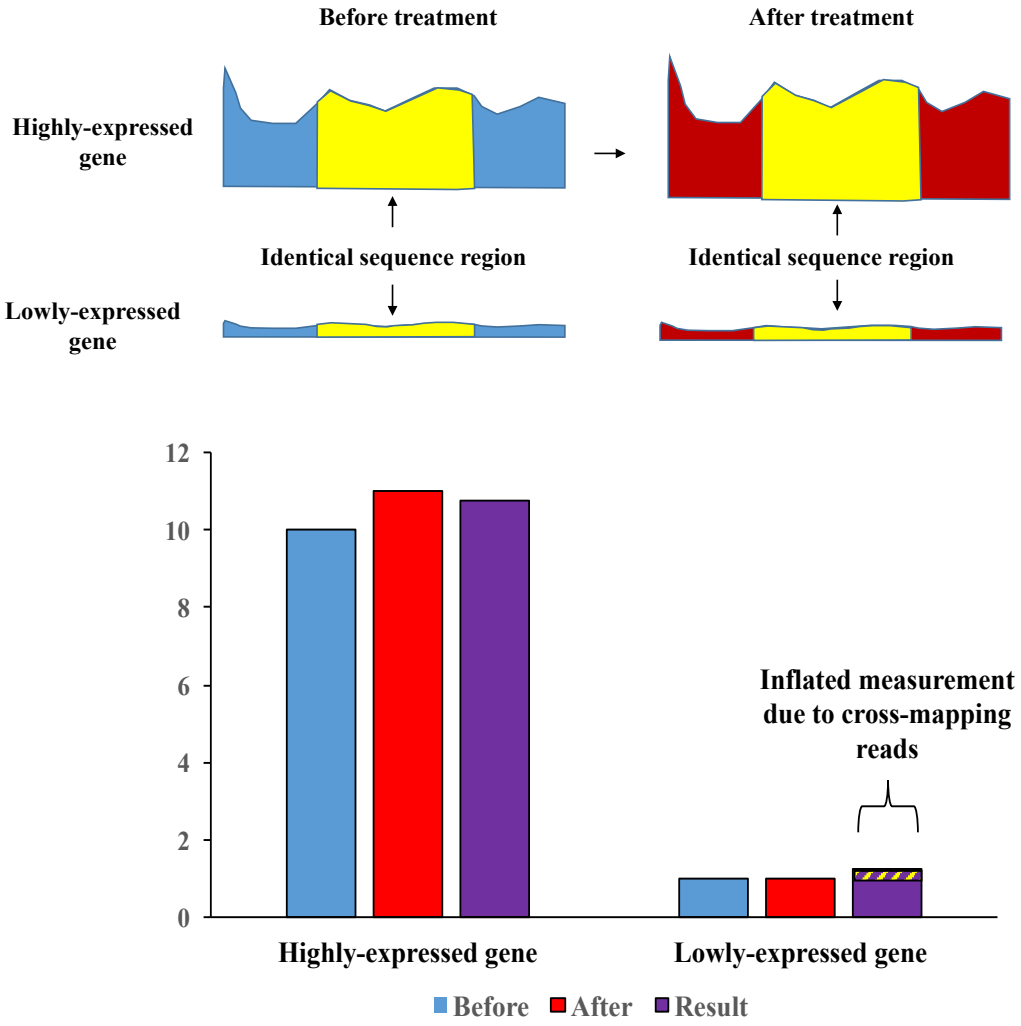


Figure 5.1 Multi-mapping reads can distort differential expression estimates. A highly expressed gene that undergoes a very small change in expression due to chemical or biological treatment may lead to inflated read counts for another gene. If the highly expressed gene shares a common sequence segment with a lowly expressed gene (yellow region in the top panel), then reads originating from the highly expressed gene may be cross-mapped to the lowly expressed gene in the post-treatment sample, giving the false impression that the lowly expressed gene had a drastic increase in activity due to the experimental treatment. Top panel indicates actual origins of various reads, not the inferred (i.e. mapped) read densities. The blue bars in the histogram represents that 'actual' expression level of each gene before the treatment, the red bars represent the 'actual' expression level of each gene after the treatment and the purple bars represent the computationally estimated expression levels after the treatment if multiple-alignments are permitted in the analysis. The highlighted area in the rightmost purple bar indicates reads originating from the highly expressed gene that are erroneously assigned to the

lowly expressed gene due to a common sequence segment (in this example, illustrating a ‘false’ increase in expression of 25%).

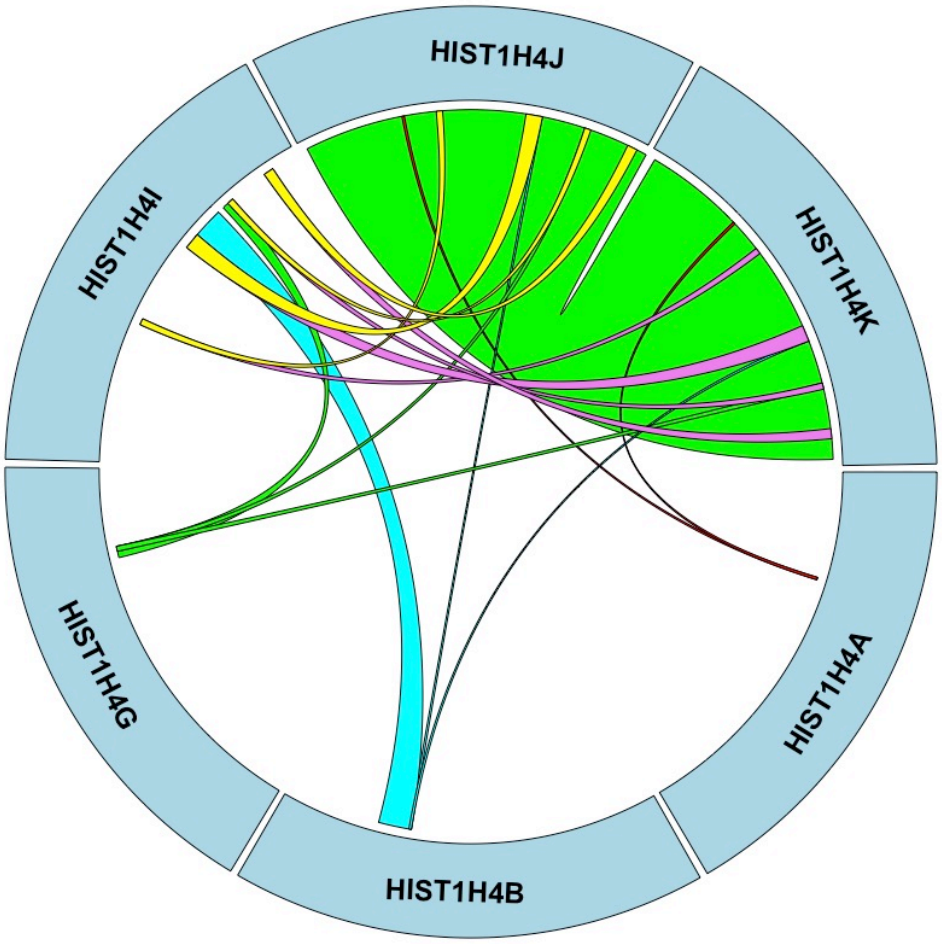


Figure 5.2 Circos plot of multi-mapping between human histone transcripts. Data from the 30-nt human CDS kmer analysis were extracted to illustrate the potential levels of inter-genic multi-mapping between six different histone genes. Each blue bar represents a different transcript, with the clockwise most portion of the bar representing the first nucleotide position in the given CDS. The various intersecting swaths indicate common sequence regions where 30-nt RPFs could be multi-mapped. Hish1h4j and Hish1h4k have more than 90% sequence similarity, while the other four selected histone ORFs have between 6 and 15% sequence similarity.

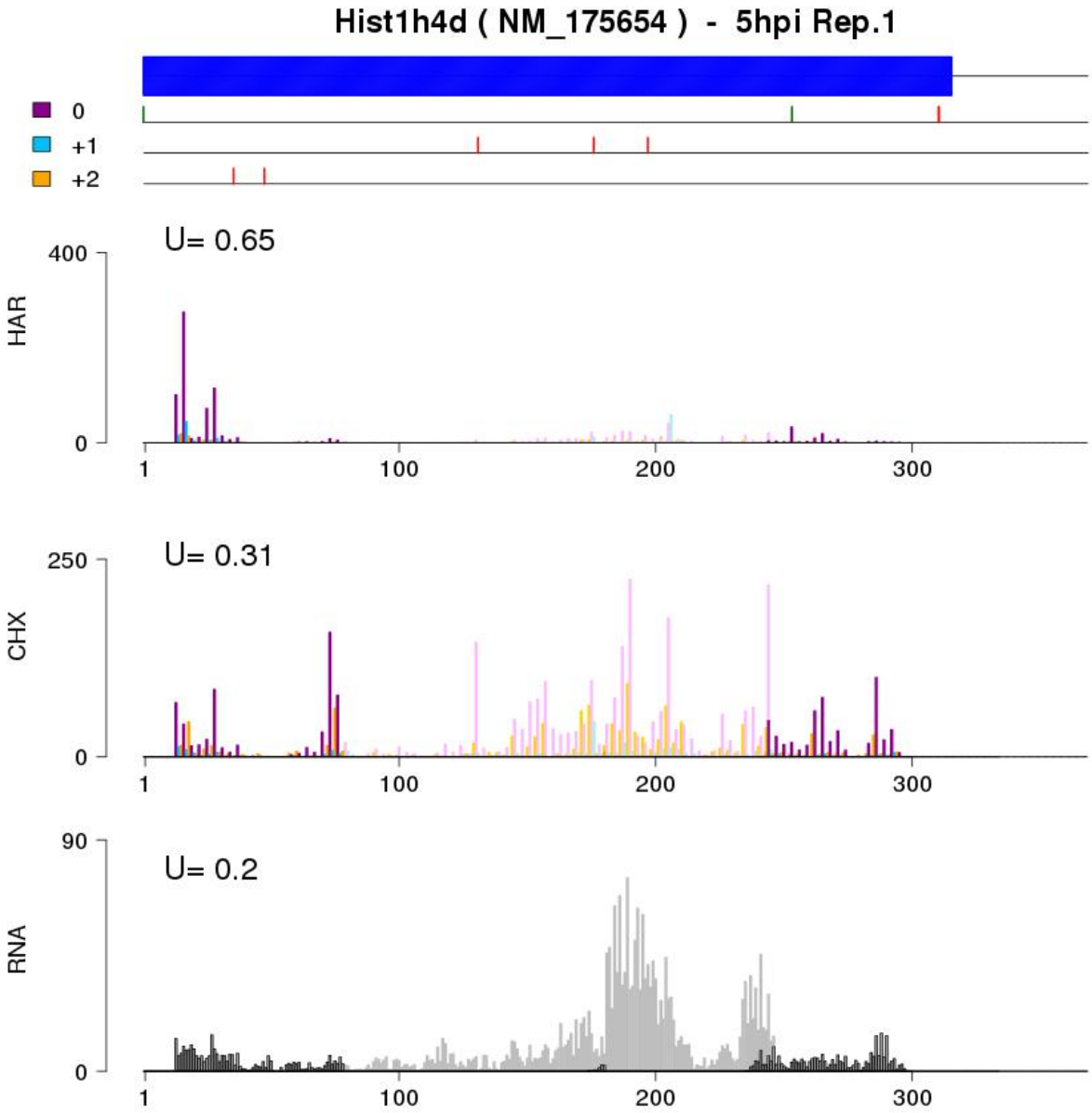


Figure 5.3 Unique and multi-mapping reads for Hist1h4d. Modified RiboSeq plot showing reads that can map uniquely to the Hist1h4d transcript in MHV-infected murine cells (dataset from Chapter 3). The light colours (from nt 80 to 240) indicate reads that could be mapped to more than one transcript, while the dark reads indicate uniquely mapped reads. The 'U-scores' indicate the proportion of reads in each plot that were uniquely mapped – for example, the bottom right panel indicates that only 20% of our RNASeq reads could be mapped uniquely to the Hist1h4d transcript while the remaining 80% of RNASeq reads that could be mapped to this transcript also had equal quality mappings on other transcripts.

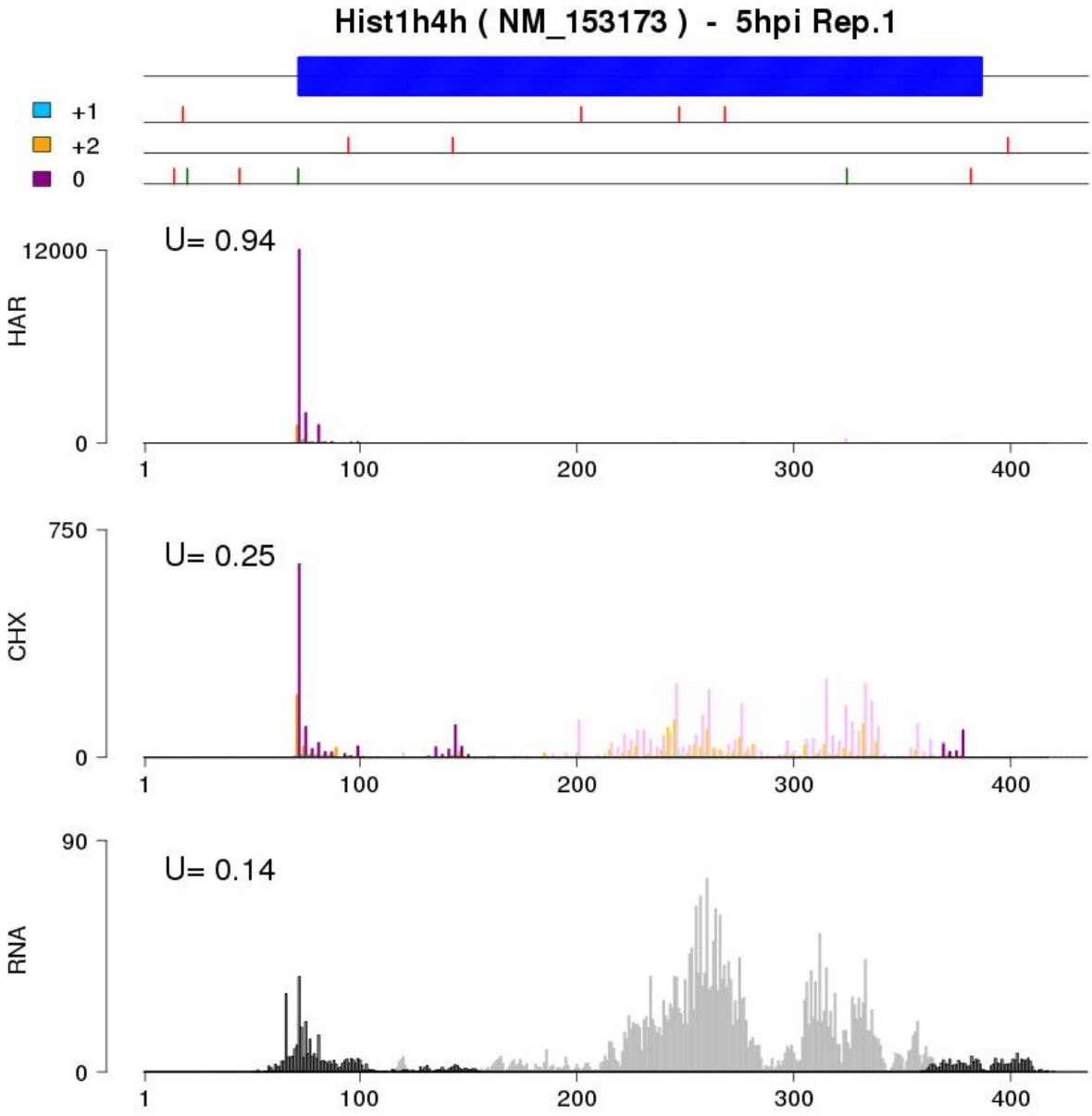


Figure 5.4 Unique and multi-mapping reads for Hist1h4h. Modified RiboSeq plot for Hist1h4h, showing uniquely mapping and multi-mapping reads for Hist1h4h in MHV-infected murine cells (dataset from Chapter 3). Light and dark colouring is as described in Figure 5.3. Some of the peaks from multi-mapping reads in Figure 5.3 are visible in this plot as well, due to regions of histone sequence similarity between Hist1h4d and Hist1h4h.

Publication	Read Alignment Algorithm Parameters
Ingolia NT, 2011	Alignments were accepted with up to two mismatches, and multiple alignments were allowed for a single sequence, but alignments with fewer mismatches were preferred
Lee S, 2012	One mismatch was allowed in all mappings; in cases of multiple mapping, mismatched positions were not used if a perfect match existed. Reads mapped more than 100 times were discarded to remove poly-A–derived reads
Reid DW, 2012	Reads from mRNA and ribosome footprinting libraries were then remapped to this subset of mRNAs, allowing each read to map to two locations, allowing two mismatches in a 25-nt seed region, and enabling the best and strata options
Liu B, 2013	The trimmed reads were aligned to the reference by SOAP 2.0 allowing up to 2 mismatches and all multiple equal best hits were retained
Stumpf CR, 2013	Clipped reads were mapped to the human genome assembly (NCBI build 36) with Burrows-Wheeler Alignment (BWA; ver. 0.5.9-r16, default parameters)
Hurt JA, 2013	RNA-seq and ribosome footprint read mapping was performed using Tophat v.1.4.0 allowing 2 mismatches but disallowing splice site mismatches and novel introns (--solexa1.3-quals --splice-mismatches 0 --min-intron-length 10 --max-intron-length 1000000 --min-isoform-fraction 0.0 --no-novel-juncs).
Crappe J, 2014	for the custom DB creation multi-mapping reads (up to 15 locations) are additionally considered
Guo JU, 2014	mapped reads to the genome using Bowtie in single-end mode, allowing ≤2 mismatches
Reid DW, 2014	Reads were mapped to an index of mouse RefSeq mRNAs (longest coding sequence for each mRNA only) using Bowtie 1.0 with a 20 nt seed region, and allowing for one mismatch, and reporting the two best locations for each read. Reads with more than 4 valid mappings were discarded
Cenik C, 2015	tophat2 -p 4 --no-discordant --library-type fr-firststrand --b2-sensitive to hg19
Grow EJ, 2015	Mapping was performed using an established pipeline previously described (Ingolia et al 2011)
Rutkowski AJ, 2015	All optimal Bowtie alignments with at most two mismatches were used
Sidrauski C, 2015	Sequencing libraries were generated as described in Ingolia et al., 2012

Table 5.1 Ribosome profiling reports that allowed multi-mapping RPFs. A selection of recent ribosome profiling experiments in murine or human cell lines that permitted some level of multiple read alignments in their read mapping pipelines (according to information in their respective methodology sections). The second column contains quotes from the method sections describing the alignment parameters used in a given study.

This analysis used the default TopHat parameters for multi-mapping reads (permitting up to 200 secondary alignments), which would likely inflate the translation efficiency calculations and potentially distort the correlation coefficient estimates. RPFdb, a recently activated online database of published ribosome profiling studies and datasets, was generated by downloading data from GEO repositories and mapping reads using STAR, allowing for multiple alignments (Xie et al., 2016). The RPFdb lists highly-expressed genes from each dataset – many of these include histones with high levels of sequence similarity, suggesting that 'double-counting' may contribute to erroneous quantification of read count data in this public database.

On the other hand, discarding all multi-mapping reads can lead to a loss of large amounts of biological information. It may also preclude the study of some gene families that share high levels of sequence similarity (some of which we describe below). Studies that omit multi-mapping reads without accounting for changes in the potential mapping space can yield biased reads per kilobase per million mapped (RPKM) estimates. One report that investigated toxin induced UPR effects on translation using ribosome profiling excluded reads with more than 4 valid mappings and calculated RPKM afterwards as a proxy for translational efficiency. However, this report does not mention any corrections in the mapping space feature lengths. If a particular gene contains 50% sequence similarity with three or more other genes, then the RPKM calculation in such a situation would need to increase the resulting estimates by a factor of two in order to account for exclusion of the multi-mapping region.

In light of these observations, we first sought to identify which coding sequences in commonly studied genomes are most likely to be affected by multi-mapping reads. To this end, we developed a series of simulations that allowed for the identification of ORFs with high levels of shared sequences.

5.2 Identification of repetitive elements in the coding sequences of various genomes

We computationally surveyed the human, mouse, yeast, *E. coli* and *Arabidopsis* genomes to identify which genes were most likely to be affected by multi-mapping reads in ribosome profiling analyses. Transcript annotations were downloaded from the NCBI RefSeq database, and annotated pseudogenes and fusion genes were discarded (

Figure 5.5). For each gene, the longest, ORF-containing isoform was selected; in cases where there were multiple equal length isoforms, the transcript with the longest ORF was kept.

If there was still a tie between two isoforms of a gene at this point, the transcript that came first in alphanumeric order was arbitrarily selected for downstream analysis. Transcripts originating from genes with overlapping genomic coordinates (such as UPK3BL/NM_001114403.2 and POLR2J2/NM_032959) were retained. After curating this list of transcripts for each genome, the annotated coding sequences were extracted using BioPython (Cock et al., 2009).

A custom C++ algorithm then splits each ORF sequence into 30-nt kmers, starting at the annotated start codon and ending 30-nt upstream of the annotated stop codon (such that that a 100-nt ORF will produce 70 kmers). The kmer sequences are then loaded into a SQlite sorted and tabulated (

Figure 5.6). The algorithm then determines how many non-unique kmers occur in each ORF. The database can then be quickly queried by kmer sequence to determine how many and which genes would be affected by non-specific mapping of that kmer.

After analyzing each 30-nt kmer in all of the ORFs, a kmer-mapping statistic (kms) was calculated for each gene. This statistic sums the number of potential alignments (M_i) for each kmer originating from a given ORF and divides the total by the ORF length minus 30 (L).

$$kms = \sum_{i=1}^{L} M_i / L$$

A ORF that only contains unique sequences would have a kms value of 1, while a ORF that is 90% identical to two other genes would have a kms value of 2.8. After calculating kms values for all of the ORFs in a given species SQLite database, the genes were sorted by kms value (in descending order) and grouped by gene families.

The murine CDSome analysis revealed 279 genes for which no kmer could be uniquely assigned to a given gene, and 828 genes for which fewer than 30% of the ORF consisted of unique 30-nt kmers (out of a total of 20427 genes). These ‘non-unique’ genes were particularly enriched for vomeronasal receptors, histones, dynein light chains, protocadherins, major urinary proteins, olfactory receptors, and keratin associated proteins (Table 5.2). Several immune system genes, including defensins, interferon precursors and histocompatibility antigen, were found to have low levels of unique kmers.

Querying the 18878 CDS elements in the human SQLite kmer database identified 173 genes for which there were no unique 30-nt sequences, and 538 for which less than 30% were unique. Non-unique genes were enriched for ubiquitin carboxyl-terminal hydrolase 17-like

proteins, histones (Figure 5.2), cancer/testis antigens, proline-rich proteins, serine-threonine kinases, nuclear-pore complex interacting proteins, olfactory receptors and zinc finger proteins (Table 5.3). Immune system genes were again found in this pool, including immunoglobulins, defensins, C-C motif chemokine precursors, complement component, G antigens and

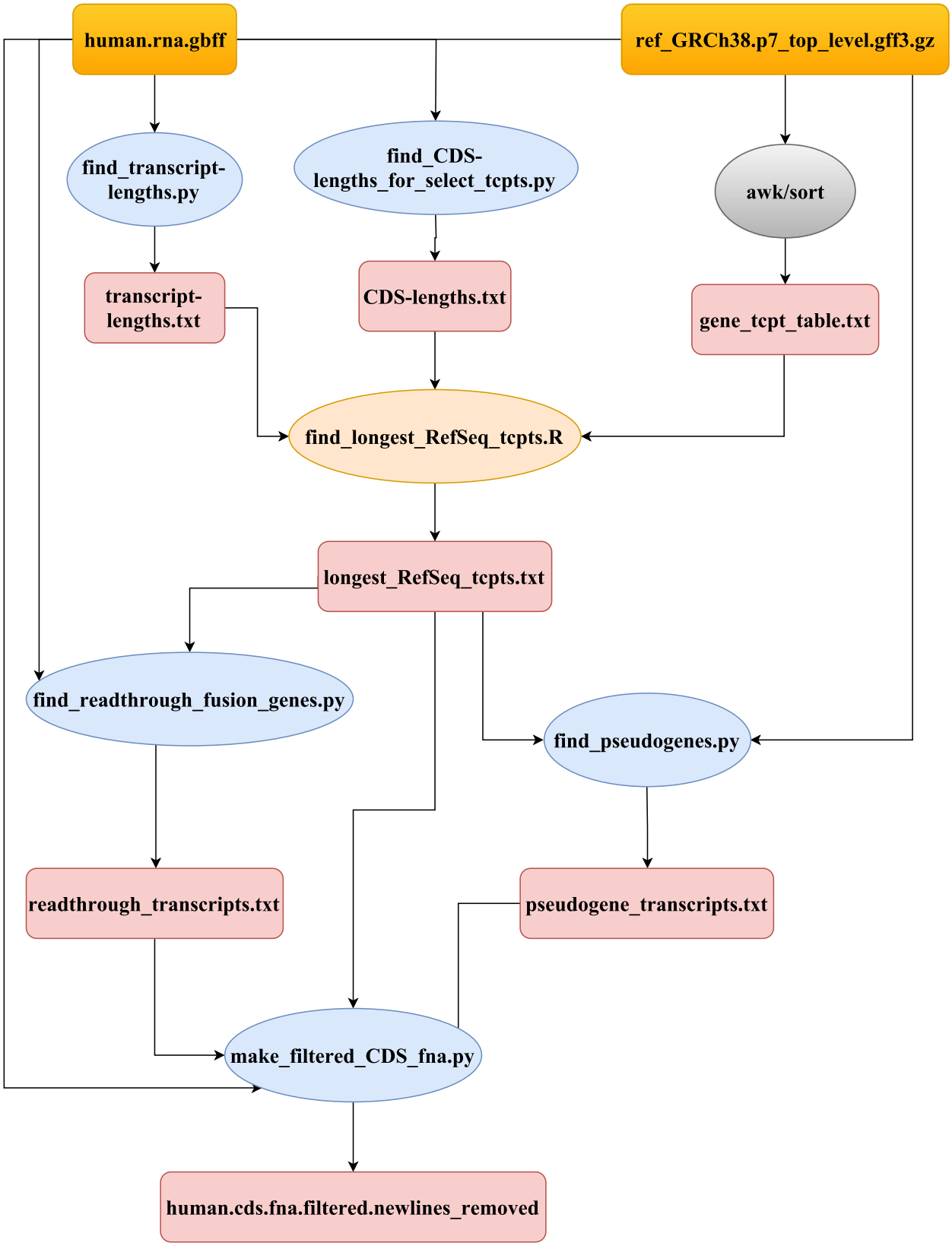

human.rna.gbff
ref_GRCh38.p7_top_level.gff3.gz
find_transcript-lengths.py
find_CDS-lengths_for_select_tcpts.py
awk/sort
transcript-lengths.txt
CDS-lengths.txt
gene_tcpt_table.txt
find_longest_RefSeq_tcpts.R
longest_RefSeq_tcpts.txt
find_readthrough_fusion_genes.py
find_pseudogenes.py
readthrough_transcripts.txt
pseudogene_transcripts.txt
make_filtered_CDS_fna.py
human.cds.fna.filtered.newlines_removed

Figure 5.5 Algorithm flowchart for kmer analysis pre-processing. In order to identify multi-mapping kmers, the first part of the investigation involved obtaining a transcriptome reference assembly and processing it to remove features that may bias the downstream analyses. Multiple isoforms of the same gene with shared exons would lead to double-counting and false calls of multi-mapping, so it was necessary to select one transcript sequence per gene. The curated NCBI assembly includes examples of pseudogenes, fusion genes and translational or transcriptional readthrough gene sequences that require filtering lest they bias the multi-mapping kmers analysis. After downloading the NCBI human Refseq transcriptome assembly, a series of biopython, R scripts and bash shell commands were used to generate a filtered list of longest protein coding transcripts for each gene. Afterwards, two biopython scripts were used to filter out pseudogenes or fusion genes. For example, transcript NM_001204088.1 is generated by a locus that represents naturally occurring read-through transcription between two neighbouring genes (Minos1, Nbl1) and yields a Minos1-Nbl1 fusion protein. To avoid double-counting kmers originating from this fusion transcript and the transcripts encoded by the separate Minos1 and Nbl1 genes, this extended transcript was filtered out from the analysis. The pipeline retains transcripts encoded by distinct genes with overlapping genomic coordinates such as UPK3BL (NM_001114403.2) and POLR2J2 (NM_032959). These elements have high levels of sequence similarity due to their shared positions on chromosome 7, but are functionally separate elements. After compiling a list of filtered transcripts, a final biopython script was used to parse the original transcriptome Genbank file and compile a single text file contain the CDS nucleotide sequences for all of the selected transcripts.

In the diagram above, ovals denote scripts or shell commands while rectangles represent data files. Biopython scripts are labeled in light blue, R scripts are in orange, bash shell commands are in grey, text data files are in red, and data files downloaded from the NCBI FTP site are in gold.

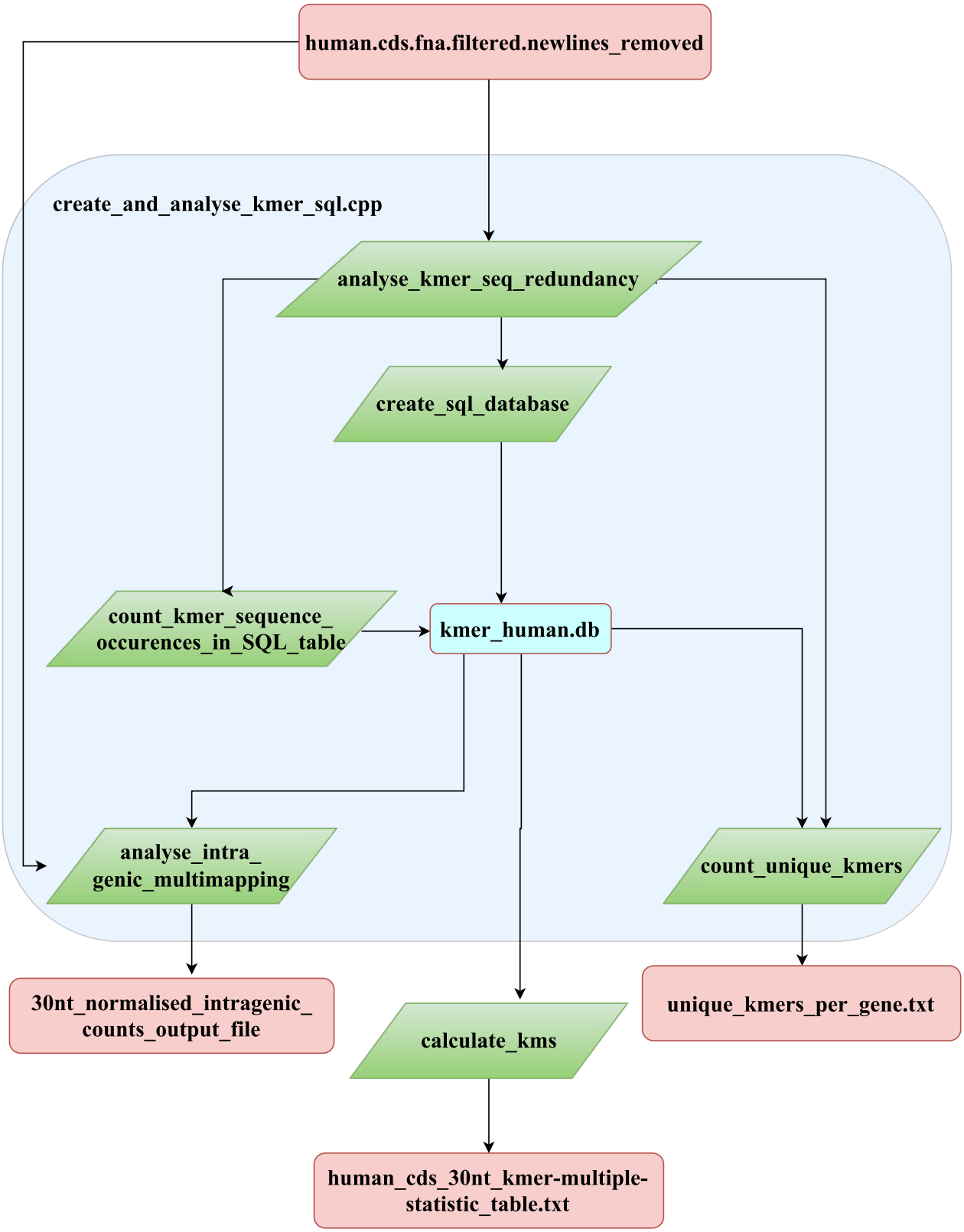

human.cds.fna.filtered.newlines_removed
create_and_analyse_kmer_sql.cpp
analyse_kmer_seq_redundancy
create_sql_database
count_kmer_sequence_
occurences_in_SQL_table
kmer_human.db
analyse_intra_
genic_multimapping
count_unique_kmers
30nt_normalised_intragenic_
counts_output_file
unique_kmers_per_gene.txt
calculate_kms
human_cds_30nt_kmer-multiple-
statistic_table.txt

Figure 5.6 Algorithm flowchart for kmer analysis. After the RefSeq transcriptome files are processed (as described in

Figure 5.5), the resulting CDS sequence file is utilized for the kmer multi-mapping simulation. This central purpose of the pipeline is to create an SQLite database that allows for quick querying of kmer counts and locations. After the CDS sequence text file is loaded, each CDS sequence is broken into a series of 30-nt kmers, which are inserted into the SQLite database along with information on their transcript source and position. Afterwards, the occurrence of each individual kmer is tabulated to identify unique and multi-mapping kmers. This information is then used to identify which genes have high-levels of uniquely mapping kmers or which have high potential for intra-genic multi-mapping. A kmer-mapping statistic (kms) is also calculated for each gene, which sums the number of potential alignments for kmer originating from a given ORF and divides the total by the ORF length.

In the diagram above, ovals denote scripts or shell commands while rectangles represent data files. Green rhomboids denote SQLite commands. Text data files are in red while SQLite database files are in cyan.

melanoma antigens. Heat shock protein HSPA1A and several TP53-target gene 3 proteins exhibited high levels of multi-mapping kmers.

The *S. cerevisiae* kmer analysis returned 87 genes for which no kmer could be uniquely assigned to a given gene, and 205 genes for which less than 30% of the ORF consisted of unique 30-nt kmers, out of a total of 5236 genes (Table 5.4). Many of these sequences originate from yeast retrotransposons that only vary slightly in sequence. The *Arabidopsis* CDSome contained 2792 ORFs which consisted solely of multi-mapping kmers and 6858 genes for which less than 30% of the ORF consisted of unique 30-nt kmers, out of a total of 31078 ORFs (Table 5.5). These genes were enriched for gametogenesis-related family proteins as well as the cysteine rich, 45-48 amino acid long antimicrobial peptides known as thionins. A large number of uncharacterized proteins were found in this set as well, which may partially be attributed to the difficulty of studying genetic elements with such high levels of non-unique sequence identities (though some of these elements may also be pseudogenes). Not surprisingly, the smaller *E. coli* genome contains only 6 genes for which no kmer could be uniquely assigned to a given gene, and 19 genes for which less than 30% of the ORF consisted of unique 30-nt kmers, out of 4032 ORFs (Table 5.6).

We sought to determine if any genes had high levels of intragenic multimapping (i.e. kmers that could be aligned to multiple distinct locations within the same given ORF). The murine polyubiquitin-C (*UBC*) transcript ORF (NM_019639.4) had 90.9% intragenic multi-mapping, meaning that less then 10% of reads from this ORF could be uniquely mapped when restricting the analysis just to this gene. The murine polyubiquitin C gene (*UBC*) is one of the two stress-regulated polyubiquitin genes. It plays a key role in sustaining the heat-shock response by maintaining cellular ubiquitin levels under stress conditions (Wiborg et al., 1985; Ryu et al., 2007). The resulting protein provides extra ubiquitin needed to tag damaged or unfolded proteins for degradation (Tsirigotis et al., 2001). The *UBC*-encoded transcript (NM_019639.4) has a repetitive nucleotide sequence that encodes nine adjacent ubiquitin domains, followed by a distinct C-terminal tail (Figure 5.7). Analyses that only utilize uniquely mapped reads are likely to miss changes in the expression of this gene, while analyses that indiscriminately retain multiple alignments will result in greatly inflated expression levels relative to other genes. Murine hornerin (74.3%, NM_133698.2), keratin associated protein 5-3 (70%, NM_023860.1), and ubiquitin B (69.4%, NM_011664.4) also had high levels of intragenic sequence redundancy. The ubiquitin genes and neuroblastoma breakpoint family member 20 (97.9%, NM_001278267.1) were similarly identified in the human CDSome analysis.

Gene Family	Number of genes with kms >= 4.0
vomeronasal 1/2 receptor	104
uncharacterized protein LOC	69
histone cluster 1/2/3/4 member	64
protocadherin alpha/gamma precursor	35
germ cell-less homolog family member	17
interferon alpha/zeta precursor	15
major urinary protein precursor	15
reproductive homeobox	14
spermatogenesis associated glutamate (E)-rich protein	13
KRAB box and zinc finger, C2H2 type domain containing	12
ovary testis transcribed	11
alpha takusan-like	10
alpha-defensin precursors	10
zinc finger protein	10
C-C motif chemokine precursor	8
preferentially expressed antigen in melanoma-like family member	8
UDP-glucuronosyltransferase precursor	8
late cornified envelope protein family member	6
RNA and export factor-binding protein 2-like	6
synovial sarcoma, X member B, breakpoint	6
serine/threonine kinase-like	5
histocompatibility antigen chain precursor	4
spermiogenesis specific transcript on the Y	4
amylase 2a precursors	3
novel KRAB box and zinc finger, C2H2 type domain containing protein	3
Slx-like 1	3
eukaryotic translation initiation factor 1A-like	2
keratin associated protein family member	2

Table 5.2 Murine gene families with high levels of multi-mapping reads. After analysing each 30-nt kmer in all of the murine ORFs (as described in

Figure 5.5 and

Figure 5.6), a kmer-mapping statistic (kms) was calculated for each gene. This statistic sums the number of potential alignments for each kmer originating from a given ORF and divides the total by the ORF length. A ORF that only contains unique sequences would have a kms value of 1, while a ORF that is 90% identical to two other genes would have a kms value of 2.8. After calculating kms values for all of the mouse ORFs in our SQLite database, the genes were sorted by kms value (in descending order) and grouped by gene families. The list above shows gene families for which one or more ORFs had a kms value greater than or equal to 4

(meaning that on average, each 30-nt kmer originating from a given ORF could be mapped to three other ORFs).

Gene Family	Number of genes with kms >= 4.0
zinc finger protein	43
ubiquitin carboxyl-terminal hydrolase 17-like	25
nuclear pore complex-interacting protein family member	15
cancer/testis antigen 47A/B	13
G antigen	12
POTE ankyrin domain family member	12
TBC1 domain family member	11
golgin subfamily A member	10
cancer/testis antigen 45	9
neuroblastoma breakpoint family member	9
PRAME family member	8
testis-specific Y-encoded protein	8
uncharacterized protein	8
RANBP2-like and GRIP domain-containing protein	7
choriogonadotropin subunit beta precursors	6
proline-rich protein 20	5
spermatogenesis-associated protein	5
forkhead box protein D4-like 3	4
variable charge X-linked protein	4
alpha-amylase 1 precursor	3
speedy protein	3
basic salivary proline-rich preprotein	2
mucin precursor	2
peptidyl-prolyl cis-trans isomerase A-like	2
COBW domain-containing protein	1
keratin-associated protein	1
killer cell immunoglobulin-like receptor precursor	1
TATA-box-binding protein	1

Table 5.3 Human gene families with high levels of multi-mapping reads. A list of human gene families for which one or more genes had a kms value greater than or equal to 4.0 (for 30-nt ORF kmers).

Gene Family	Number of genes with kms >= 4.0
gag protein	42
gag-pol fusion protein	40
seripauperin	14
Y element ATP-dependent helicase	9
asparaginase	4
Cos2/Cos3 protein	2
Flo11 protein	1

Table 5.4 Yeast gene families with high levels of multi-mapping reads. A list of *Saccharomyces cerevisiae* gene families for which one or more genes had a kms value greater than or equal to 4.0 (for 30-nt ORF kmers). The yeast RefSeq database contains many retrotransposon elements that encode highly-similar gag proteins.

Gene Family	Number of genes with kms >= 4.0
uncharacterized protein	61
ECA1 gametogenesis related family protein	31
plant thionin family protein	12
phosphoglycerate mutase-like protein	8
nuclear transcription factor Y subunit B-1	8
UDP-3-O N-acetylglucosamine deacetylase	8
cysteine-rich repeat secretory protein	8
WD40 repeat protein MUCILAGE-MODIFIED 1	7
sirtuin 2	7
actin bundler LIM family protein WLIM2b	7
transcription factor TGA1	6
transcription factor BIM1	6
Shaggy-related protein kinase kappa	6
ribosomal protein S5 domain 2-like superfamily protein	6
putative transcription factor PosF21	6
putative RAB GTPase activator protein	6
peroxisomal (S)-2-hydroxy-acid oxidase GLO1	6
otubain-like histone deubiquitinase 1	6
glyoxalase I homolog GLX1	6
cysteine synthase D2	6
UBP1-associated protein 2A	5
transcription factor TGA6	5
sterol 4-alpha-methyl-oxidase 2-2	5
RNA recognition motif-containing protein	5
putative plastid-lipid-associated protein	5
protein kinase family protein	5
protein brassinazole-resistant 2	5
protein AGAMOUS-like 42	5
post-illumination chlorophyll fluorescence increase protein	5
PB1_UP2 domain-containing protein	5
lysophosphatidyl acyltransferase 5	5
histone-lysine N-methyltransferase SUVR2	5
E3 SUMO-protein ligase SIZ1	5
carbonic anhydrase 2	5
BTB/POZ domain-containing protein	5
ACT domain-containing protein 3	5

Table 5.5 Arabidopsis gene families with high levels of multi-mapping reads. A list of *Arabidopsis* gene families for which one or more genes had a kms value greater than or equal to 4.0 (for 30-nt ORF kmers).

Gene Family	Number of genes with kms >= 2.0
toxic membrane polypeptide, small	3
Rhs protein with putative neighboring cell growth inhibitor	2
3-hydroxypropionic acid resistance peptide	1
DLP12 prophage; putative lipoprotein	1
putative Rz1-like lipoprotein, Qin prophage	1
Rac prophage; putative lipoprotein	1

Table 5.6 *E. coli* gene families with high levels of multi-mapping reads. A list of *E. coli* gene families for which one or more genes had a kms value greater than or equal to 4.0 (for 30-nt ORF kmers).

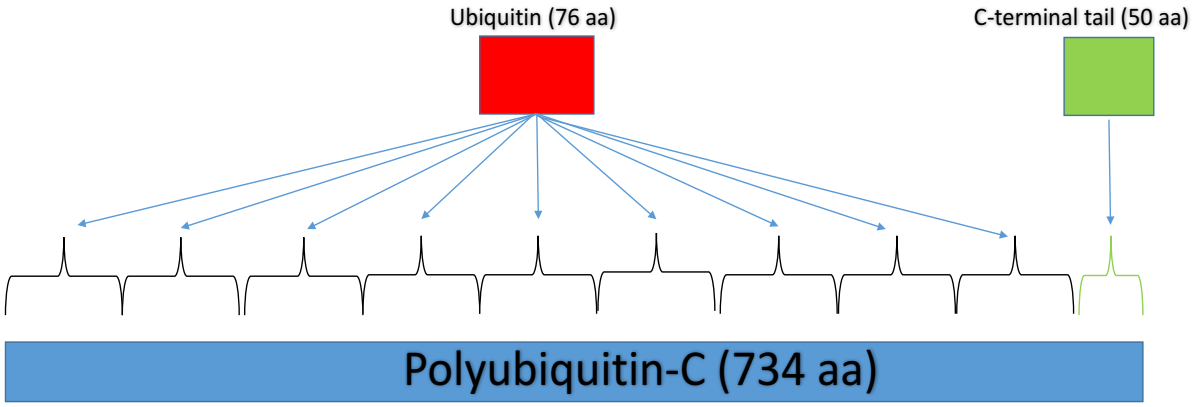


Figure 5.7 Murine polyubiquitin C gene has high levels of intra-genic multi-mapping. The murine polyubiquitin C gene (*UBC*) is one of the two stress-regulated polyubiquitin genes. It plays a key role in sustaining the heat-shock response by maintaining cellular ubiquitin levels under stress conditions (Wiborg et al., 1985; Ryu et al., 2007). The resulting protein provides extra ubiquitin needed to tag damaged or unfolded proteins for degradation (Tsirigotis et al., 2001). The *UBC*-encoded transcript (NM_019639.4) has a repetitive nucleotide sequence that encodes nine adjacent ubiquitin domains, followed by a distinct C-terminal tail. Analyses that only utilized uniquely mapped reads are likely to miss changes in the expression of this gene, while analyses that indiscriminately retain multiple alignments will result in inflated expression levels. The red box denotes the 76 amino acid ubiquitin domains, while the green box denotes the 50 amino acid C-terminal tail with a unique sequence.

5.3 Strategies for handling multi-mapping reads

No tools exist specifically for the handling of multi-mapping reads in the context of RiboSeq experiments, though several approaches developed for RNASeq experiments can be readily incorporated into profiling computational pipelines. After mapping our murine RPF data from Chapter 3 against host rRNA sequences using Bowtie, we mapped non-rRNA reads with RSEM, setting the parameters *-p 8 --seed-length 25 --bowtie-n 2 --bowtie-m 200 --sampling-for-bam --phred33-quals --estimate-rspd --append-names --output-genome-bam --sort-bam-by-coordinate* (Li et al., 2009; Langmead, 2010). This aligns reads to the murine reference transcriptome, using Phred+33 format quality scores and allowing for 2 mismatches in the seed region and suppressing all alignments for a read if more than 200 valid alignments exist. The *--estimate-rspd* parameter causes the algorithm to partition each transcript into 20 bins and count the frequency of reads start at each bin. This allows the algorithm to estimate the read start position distribution (RSPD) from the data (to assess if there are positional biases in the data) instead of assuming a uniform RSPD.

The *--output-genome-bam* and *--sort-bam-by-coordinate* arguments cause the program to generate an index-sorted BAM file with alignments mapped to genomic coordinates and annotated with their posterior probabilities. When *--sampling-for-bam* is activated, one alignment is sampled according to the posterior probabilities and the remaining potential alignments are discarded. This allows for the use of the data in differential expression packages that do not account for multi-mapping posterior probabilities. The resulting genomic BAM file was then tabulated with HTSeq-count, restricting for alignments falling within annotated coding sequences (Anders et al., 2015).

The new count files were then used in a similar fashion as in Chapter 3 to perform differential translation analyses with Xtail (Xiao et al., 2016). For the majority of genes, reads counts were similar regardless of whether unique or multi-mapping alignments were used. However, a sizeable minority had large changes in read counts. For the 5 h p.i. analyses with four mock and two infected replicates, 87% of the 335 genes previously identified as differentially expressed were similarly labelled in the new analysis. However, 23 genes were identified as no longer being significantly affected by MHV infection, while a new set of 19 genes was identified as being differentially translated (Table 5.7). Many histones appear in both categories, as would be expected given the potential for multi-mapping reads between various histone genes to affect inter-replicate read count variability.

No Longer DE	New DE
Rpl29	Rpl31-ps12
Hist1h4k	Hist2h3c2
Als2	Hist1h2bj
Hist1h2bn	Mfap1b
Hist1h2ai	Hist1h4i
Hist1h2bh	Rsph3a
2200002D01Rik	Hist1h2be
Cks1b	Hist1h3a
Fxyd5	Ddhd2
Hist1h2bg	H2afx
Hist1h2bf	Rps15a
Hist1h2ab	Slc9b1
Hist1h3e	Man1a2
Hist1h4f	Hist1h3c
Hist1h2bp	Abcb1b
Pyroxd1	2610524H06Rik
Ube2f	Hist1h2bb
Hist2h3b	EU599041
2810417H13Rik	Chtf8
Hist2h2ac	
Rpl31	
Hoxa7	
Tmem192	

Table 5.7 Changes in differential translation analyses when incorporating muli-mapping reads. 19 new genes were identified as differentially expressed in our MHV-infected murine cell datasets (as describe in Chapter 3) when multi-mapping reads were retained using RSEM. These analyses were repeated using Xtail and similar cut-offs (FDR < 0.05 and log2 fold change greater than 1 or less than -1). 23 genes were no longer detected as differentially expressed.

The original differential expression analysis (Chapter 3) discarded multi-mapping reads when HTSeq-count was invoked, as this discards reads for which the 'NH:i' tag has a value greater than zero. Unlike TopHat, RSEM assesses the potential for a multi-mapping read to originate from a particular locus based off the alignments of other reads. One potential disadvantage of RSEM is that it does not have built-in TopHat compatibility and therefore cannot perform *de novo* splice site identification. For certain research objectives, it may be necessary to run separate pipelines involving RSEM and TopHat and to cross-reference the outputs.

5.4 Discussion

To our knowledge, this is the first study to explicitly investigate the impact of multi-mapping reads on ribosome profiling analyses. We identified several sets of gene families that are likely to be affected by ambiguously mapping reads, including histones, olfactory receptors and various immune receptor genes. We re-analysed our datasets from Chapter 3 using a probabilistic multi-alignment allocation algorithm (RSEM) and show how retention of multi-mapping reads affects visualization and differential expression analyses of RPFs.

Many ribosome profiling studies have either discarded multi-mapping reads or retained them without sorting through the various potential alignments, leading to 'double-counting' artefacts. Several published protocols do not include steps for proper handling of such reads (Table 5.8). Studies that choose to discard multi-mapping reads without adjusting for the reduction in mapping space will result in distorted RPKM estimates. The magnitude of the multi-mapping problem is increased by allowing for up to two mismatches in the RPF alignment, as is commonly done in practice. Paralogous gene families, low-complexity sequences, reference sequence errors, sequencing errors that allow for mismatches or indels in read alignments, transcripts originating from genes with overlapping genomic coordinates and repetitive protein domains all contribute to the potential for RPF mapping ambiguity. In order to reduce the potential for computational artefacts as well as maximize salvaging of biological information, we suggest future profiling studies use an algorithm such as RSEM or eXpress to handle multi-mapping reads in a rigorous fashion (Chung et al., 2011; Roberts and Pachter, 2013).

One of the limitations of RSEM is that it does not allow for splice-junction detection and introduces an additional (modest) computational burden. We also show that some ORFs have no unique 30-nt sequences, making it extremely difficult to establish whether RPFs originate

Publication	Read Allignment Algorithm Parameters
Ingolia NT, 2012	tophat --no-novel-juncs --output-dir XXX_vs_genome --GTF hg19.gtf hg19 XXX_norrna.fq ; samtools view -h XXX_vs_genome/accepted_hits.bam | grep –E '(NM:i:0)|(^@)' | samtools view –S b – > XXX_vs_genome.bam
Ingolia NT, 2013	tophat --no-novel-juncs --output-dir XXX_vs_genome --GTF hg19.gtf hg19 XXX_norrna.fq ; samtools view -h XXX_vs_genome/accepted_hits.bam | grep –E '(NM:i:0)|(^@)' | samtools view –S b – > XXX_vs_genome.bam
Illumina Ribosome Profiling Kit Bioinformatics Guide	topHat --GTF /Annotation/Genes/genes.gtf --numthreads 1 --output-dir /Sequence/BowtieIndex/genome .fastq

Table 5.8 Published ribosome profiling protocols that retain multi-mapping reads. A selection of published protocols or guides that lead to retention of multiple read alignments at the mapping stage. The "NM:i:0" tag described in the first two protocols refers to the number of mismatches in a given alignment in the SAM format (i.e. 'perfect' matches). The Illumina guide refers to a bioinformatics guide that is released in association with the Illumina TruSeq (formerly ARTseq) Ribo Profile kit.

from these genes. Selectively limiting the RPF mapping space by using RNASeq data to modify alignment reference databases may aid in reducing some mapping ambiguities. This tactic could be further augmented by using longer and/or paired-end RNASeq reads, as well as potentially using alternative sequencing platforms to quantify individual isoforms and full molecules of mRNA in the future. The RSEM expectation-maximization algorithm can estimate a read start position distribution and learn fragment length distributions from the data, though in its current implementation it may not be optimised for dealing with RiboSeq data. RPFs tend to map to ORFs and exhibit three nucleotide periodicity; future work will investigate whether these features can be incorporated into the RSEM model to improve performance for these datasets.

In situations where it is known *a priori* that an investigation will not include the additional experimental or computational efforts needed to allow for analyses of genes with high levels of multi-mapping kmers, it may be optimal to discard multi-mapping reads and avoid assessing for differential expression of problematic genes. With high-dimensional data such as RiboSeq, gene by gene statistical testing is used to select genes whose translational efficiency differs across conditions (Bourgon et al., 2010). Such analyses require adjustment for multiple testing, which can result in low statistical power. Multiple testing adjustments such as the Benjamini-Hochberg procedure or Bonferroni correction provide control over the extent to which false positives occur, but such rigour decreases the power to detect true positives for differential expression. The power reduction becomes more severe as more hypotheses are tested.

As many ribosome profiling analyses lack the sensitivity to detect expression of genes prone to multi-mapping reads, it may make sense to remove sets of these genes for which the RiboSeq experiment will generate an uninformative signal. Such a filter will reduce the number of variables being tested, making the multiple testing corrections less restrictive and thus enhancing power of detection of true differential expression. The databases generated in the course of this work can aid in quickly constructing such multi-mapping gene filter lists.

Chapter Six

Study of Plant Translation During Temperature Stress

6 CHAPTER SIX: STUDY OF PLANT TRANSLATION DURING TEMPERATURE STRESS

6.1 Introduction

Though the 80S ribosome and many translation factors are conserved across *Eukarya*, there exist many kingdom-specific mechanisms of translation regulation that evolved to meet taxon-specific environmental challenges. Plants rely on photosynthesis as the major source of glucose, are poikilotherms, and are sessile, which prevents them from being able to relocate in response to various abiotic/edaphic (seasonal and diurnal changes in temperature and lighting, changes in water availability, changes in soil phosphate and nitrate concentration, etc.) and biotic stresses (e.g. herbivores).

Plants have specific epigenetic, transcriptional, translational and post-translational mechanisms that sense changes in these environmental parameters and mediate specific responses (Kumar and Wigge, 2010). Several studies in the past 15 years have used microarrays or RNASeq to interrogate changes in plant transcriptomes in response to cold or heat stress (Fowler and Thomashow, 2002; Winfield et al., 2009, 2010; Rasmussen et al., 2013). These studies revealed that plants such as *Arabidopsis* or wheat undergo a gradual modulation in abundance of several transcripts encoding cold-sensitive transcription factors in response to low temperatures that enable survival of thermal stress; rapid decreases in temperature do not allow for this response to occur. In microarray-based analysis of the *Arabidopsis* transcriptome, 24-hour cold exposure resulted in significant changes in 4% of transcripts (Lee et al., 2005), while exposure for 14 days caused 20% of transcripts to be differentially expressed (Hannah et al., 2005). Polysome profiling of *Arabidopsis* seedlings exposed to darkness at mid-day resulted in a 17% decrease in translation initiation, which was reversed within 10 minutes of re-illumination (Juntawong and Bailey-Serres, 2012). ~1600 mRNA species were found to be differentially translated in response to changes in ambient lighting, particularly those transcripts encoding protein synthesis and photosynthesis components. Other studies using polysome profiling have found that a number of other stressors can induce rapid changes in translation efficiency of select groups of transcripts; these factors include cadmium exposure, dehydration, high salinity, high temperature, hypoxia, gibberellin treatment, ozone-induced oxidative stress and sugar starvation (Branco-Price et al., 2005; Nicolaï et al., 2006; Mustroph et al., 2009; Matsuura et al., 2010; Puckette et al., 2012; Ribeiro et al., 2012). Juntawong et al. (2013) found that cold shock in *Arabdiopsis* leads to induction of a mRNA chaperone cold-shock protein 1 (CSP1) that enhances ribosome loading on select transcripts, including ribosome biogenesis

factors. Sorenson and Bailey-Serres (2014) found that the oligouridylate-binding protein 1 arrests translation during hypoxic conditions by clustering and sequestering transcripts in a quickly reversible manner.

As mentioned previously, a handful of studies have utilized ribosome profiling to study translation regulation in plants. The first published plant RiboSeq paper revealed that plant ribosomes enclose roughly the same length of mRNA as yeast and mammalian ribosomes (Liu et al., 2013b). AUG-initiated uORFs were found to repress the translation of main coding sequences in a light-dependent manner, though CUG-initiated uORFs did not have this effect. Additionally, transcripts encoding components of the chloroplast photosynthetic apparatus were translationally upregulated in response to light stimulation. Zoschke et al. (2013) utilized a variant of ribosome profiling for the study of chloroplast translation and found that ribosome occupancy of the first exon of the chloroplast *atpF* mRNA was identical for spliced and unspliced transcripts, corroborating previous data that suggested that splicing is not a requirement for translation of chloroplast mRNAs. Juntawong et al. (2014) found hypoxia-responsive mRNAs were upregulated at the level of transcript abundance but did not have a disproportionate change in RPFs, suggesting that these genes are not translationally upregulated in response to hypoxia but are primarily regulated at the transcriptional level. Additionally, the study found that 32% of translationally upregulated transcripts under hypoxic conditions had uORFs and that translationally downregulated transcripts were highly enriched for those encoding ribosomal proteins. Another study found that during chloroplast differentiation in maize, translation efficiency largely functioned to amplify transcriptional changes (Chotewutmontri and Barkan, 2016).

RiboSeq offers several advantages over previous methodologies used to study plant translation, such as a wider detection range and superior resolution than polysome profiling. This chapter discusses the use of ribosome profiling to further investigate how cold and heat stress can modulate plant translation. Plant protein synthesis (particularly in *Arabidopsis thaliana*) is briefly reviewed, followed by descriptions of RiboSeq meta-analyses and differential gene expression identification.

6.1.1 Plant translation

The mechanisms and principles of plant cytoplasmic translation are in many ways similar to those of yeast and mammals, with some notable evolutionary idiosyncrasies related to processes such as photosynthesis (Browning and Bailey-Serres, 2015; Dutt et al., 2015). Plant

cytosolic ribosomes have several unique proteins, post-translational modifications and diverse paralogs (Carroll, 2013). The 79 ribosomal protein families of the mammalian ribosome are found in plants (Wilson and Cate, 2012), but the acid stalk protein P3 is unique to plants (Chang, 2005; Carroll et al., 2008; Carroll, 2013). In mammals, each ribosomal protein is typically encoded by one gene (Sugihara et al., 2010), while plant ribosomal proteins are expressed from three or four genes each (Carroll, 2013). Several proteomic studies have confirmed the production of multiple distinct isoforms from each family of paralogous ribosomal proteins (Giavalisco et al., 2005; Carroll et al., 2008; Piques et al., 2009; Turkina et al., 2011; Hummel et al., 2012). Though there is debate about the role of heterogeneous ribosome composition, there is evidence to suggest that differential incorporation of distinct ribosomal proteins has effects on plant leaf development and in response to changes in sucrose availability (Horiguchi et al., 2011; Hummel et al., 2012). Cytosolic *Arabidopsis* ribosomes undergo coordinated cycles of phosphorylation and dephosphorylation of several serine residues between day-time and night-time that correlate with global shifts in protein synthetic activity (Turkina et al., 2011).

Plant initiation factors are primarily conserved with other eukaryotic taxonomic groups (Muench et al., 2012), though eIF(iso)4E and eIF(iso4)G are only found in plants (Hernández and Vazquez-Pianzola, 2005; Lellis et al., 2010). *Arabidopsis* has been found to have more distinct factors that interact with the C-terminal domain of poly(A)-binding proteins than other eukaryotes (Bravo et al., 2005).

While a discussion of plant transcription is outside the scope of this text, it is worth mentioning that plants possess many similar transcript quality control mechanisms to other eukaryotic systems. Several nonsense-mediated decay pathway and exon-junction complex orthologs have been found in plant genomes (Pendle et al., 2005; Arciga-Reyes et al., 2006; Park and Muench, 2007).

Many of the global translational control mechanisms characterized in metazoans have been found to occur in plants as well. The unfolded protein response that senses proteostatic stress in the endoplasmic reticulum has been shown to be important for plant homeostasis (Chakraborty et al., 2016). In plants, the IRE1 sensor also functions as a stress sensor and selective splicing factor, regulating the production of an alternative bZIP60 isoform that functions as a transcription factor to increase UPR gene expression. The equivalent of the Atf6 branch in plants consists of bZIP28 and bZIP17 cleavage, which allows part of these proteins to translocate to the nucleus and induce upregulation of UPR effectors. The third branch of the yeast/mammalian UPR, the PERK-mediated phosphorylation of eIF2α, has not been confirmed to occur in plants (Immanuel et al., 2012; Chakraborty et al., 2016). Stress granules and

processing bodies have also been found to occur in plants, with the involvement of the decapping proteins DCP-1 and DCP-2 (Xu and Chua, 2009). These structures have been shown to be vital for programmed translational repression during post-embryonic plant development (Xu and Chua, 2011). Some plant transcripts contain an internal light response element (iLRE) - an RNA sequence regulatory element that helps mediate illumination-dependent reversible translational suppression (Roy and von Arnim, 2013).

6.1.2 Heat Shock and Cold Shock Response

Changes in geographic temperature patterns have been found to affect the global distribution of plant species and may influence crop cultivation (Mittler, 2006; Kelly and Goulden, 2008; Lenoir et al., 2008). Given that the focus of this ribosome profiling study was to evaluate the effects of heat and cold stress on *Arabidopsis* translation and transcript abundances, a brief summary of the plant heat and cold shock responses will be provided. Plants are exquisitely sensitive to fluctuations in ambient temperature and can adjust their physiology in response to differences of as little as 1°C (Kumar and Wigge, 2010). Diurnal and seasonal thermocycles have been shown to entrain endogenous biological clocks, which include genetic modules that regulate the timing of plant reproduction (Sung and Amasino, 2005; Heggie and Halliday, 2005; Michael et al., 2008; Haydon et al., 2013; Song et al., 2013). Certain strains of wheat and other plants require an extended period of cold for flowering (vernalization) (Winfield et al., 2010).

Some species of plants have evolved the ability to gradually acclimate to cold or hot temperature ranges over the course of several days or weeks by increasing expression of various protective proteins and metabolites (Zhu et al., 2007). Plants from colder climates, such as winter wheat, barley, oat, rye and oilseed rape can cold-acclimate and acquire tolerance to inter-cellular ice formation in their vegetative tissues (Chinnusamy et al., 2007). However, many important crop species are cold-sensitive and are not able acclimate to cold spells nor can they tolerate the formation of ice crystals within their tissue. These include crops such as rice, maize, soybean, cotton and tomato. Extreme temperatures prompt changes in transcription, chromatin remodeling, alternative splicing, miRNA and siRNA-mediated transcript turnover and translation (Branco-Price et al., 2005; Chinnusamy et al., 2007; Filichkin et al., 2010; Kumar and Wigge, 2010; Park et al., 2012; Sunkar et al., 2012; Ambrosone et al., 2012). Several independent investigations have noted the increased transcription of a set of stress-related genes, particularly those encoding heat shock proteins, which function as protein chaperones

(Huang and Xu, 2008). Other upregulated genes include those involved in heat-induced oxidative protection, such as thioredoxin, glutathione S-transferase, dehydroascorbate reductase and cytosolic Cu/Zn-superoxide dismutase (Ferreira et al., 2006; Lee et al., 2007; Huang and Xu, 2008). In *Arabidopsis*, heat stress induces variable levels of polysome dissociation in most transcripts within 10 minutes, though a subset of transcripts are not translationally repressed (Matsuura et al., 2010). The latter group is enriched for genes involved in stress response, protein ubiquitination/degradation, histone/chromatin structure and RNA-mediated regulation of transcription. Heat shock downregulates ribosome biogenesis factors, photosynthesis components, cell wall precursor synthesis mediators, as well as lipid, amino acid and nucleotide metabolism. Many of the translationally resistant transcripts harbour a *cis*-regulatory element in the 5′ leader sequence that affects differential translation in response to heat stress (Matsuura et al., 2013). For example, the *Arabidopsis* HSP81-3 mRNA 5′ UTR harbours a 47-nt pyrimidine rich element that can promote cap-independent translation after heat shock (Matsuura et al., 2008). Unlike the heat shock response in metazoans, in plants there has been no observed change in phosphorylation of GCN2 or eIF2α following induction of stress (Matsuura et al., 2010). However, for cold shock, there is evidence of both these mechanisms being activated to induce translational repression (Echevarría-Zomeño et al., 2013). Some of the heat shock resistant transcripts contain 5′-UTRs that have relatively low requirements for cap–eIF4E interaction-dependent processes, and it has been suggested that the inhibition of cap–eIF4E interactions plays a role in the selection of transcripts for translation during heat stress (Dinkova et al., 2005; Matsuura et al., 2013). Heat shock has been documented to result in the transient sequestration of certain transcripts into stress granules in tomato, carrot and *Arabidopsis* cells (Roy and von Arnim, 2013; Maldonado-Bonilla, 2014).

Data from several studies in the past decade suggests that the UPR also plays a role in modulating translation during the heat shock response (Chakraborty et al., 2016). The *Arabidopsis* equivalent of mammalian Atf6, basic leucine zipper 28 (AtbZIP28) has been shown to promote resistance to heat stress; deletion of this gene severely affects plant phenotypes (Gao et al., 2008; Moreno and Orellana, 2011). The maize orthologue of mammalian Xbp1, ZmbZIP60 is alternatively spliced and activated during heat stress as well (Li et al., 2012c).

Unlike heat stress, cold stress tends to slow the kinetics of endogenous reactions. Cold stress results in elevated expression of cold shock proteins (CSPs) which contain ~70 amino acid nucleic acid binding cold shock domains (CSD) (Sasaki and Imai, 2012). The inducer of CBF expression 1 (ICE1) is sumoylated by SUMO E3 ligase and upregulates expression of C-

repeat binding factor (CBF), which in turn increases transcription of cold response genes in *Arabidopsis*, *Brassica*, tomato and rice (Lissarre et al., 2010). In *Arabidopsis*, cold-induced membrane rigidity triggers the diacylglyerol kinase pathway (Chinnusamy et al., 2007). Cold stress induces the accumulation of proline, which activates genes regulated by proline-responsive element (PRE) containing promoters (Satoh et al., 2002). Proline also functions as an osmo-protectant. The *Arabidopsis* genome encodes four known proteins that contain N-terminal CSDs: AtCSP1–AtCSP4 (Sasaki and Imai, 2012). A recent study extensively characterized the role of AtCSP1 and found that the stress protein associates with polysomes via an RNA-mediated interaction (Juntawong et al., 2013). AtCSP1-associated transcripts contain GC-rich 5′ leaders and are enriched for cellular respiration, mRNA binding and translation (particularly ribosome biogenesis) related ontologies. After induction of cold stress, AtCSP1 levels increase and correlate with an enhanced translation of ribosomal protein transcripts. It has therefore been proposed that a key function of AtCSP1 and other cold shock response factors in cold-tolerant plants is to facilitate translation of certain cellular components during low temperatures.

6.2 Temperature-induces changes in *Arabidopsis* gene expression

Two week old *Arabidopsis thaliana* seedlings were grown in ½ x MS media and kept at 4, 17 or 28 °C. RiboSeq and RNASeq libraries were prepared by Dr Betty Chung (Dept. of Plant Sciences) as in Section 2.2 and libraries were multiplexed and sequenced on an Illumina HiSeq platform at the Beijing Genomics Institute (with only one sample per temperature condition). Reads were aligned to the TAIR10 *Arabidopsis thaliana* transcriptome assembly from The Arabidopsis Information Resource (TAIR) database using Bowtie (Langmead, 2010; Lamesch et al., 2012).

6.2.1 Differential gene expression

Two sets of differential translation analyses were performed using Xtail (Xiao et al., 2016). The 17 °C RiboSeq/RNASeq libraries were compared with the 4 °C libraries to the effects of cold stress on translation. The 17 °C libraries were separately compared against the 28 °C treated samples to study the effect of heat stress on ribosome activity. The expression results were filtered using a log2 fold change greater than 1 or less than -1, and a

False Discovery Rate value less than 0.05. In the cold stress treated sample, 47 transcripts translationally downregulated and 62 were translationally upregulated according to these criteria (Figure 6.1,

Figure 6.2). Consistent with previous studies (Juntawong et al., 2013), the translationally upregulated set was enriched for genes involved in ribosome biogenesis (Table 6.1). Organonitrogen compound response genes were upregulated both at the translational and transcriptional level, suggesting that for this pathway the increase in translational efficiency serves to amplify the genetic signal in response to cold stress (Table 6.2). Poly(A) tail shortening genes were downregulated at both the transcript abundance and translational efficiency levels, possibly as a part of a mechanism to transiently inhibit the rate of transcript degradation.

In the 28 °C differential expression analysis, 342 genes were significantly upregulated and 539 were significantly downregulated at the level of translation (Figure 6.3, Figure 6.4). In agreement with previous reports (Matsuura et al., 2010), genes that encode components of the photosynthetic apparatus and translation machinery were downregulated translationally, while those corresponding to oxidative stress response and heat shock proteins were upregulated (Table 6.3, Table 6.4).

6.2.2 Effects of GC content, CDS length and UTR length on gene expression

As mentioned earlier, there is considerable evidence that various intrinsic features of plant mRNAs correlate with transcript stability and translation efficiency (Yángüez et al., 2013). The average GC content of plant 5′ UTRs is significantly lower than that of mammalian leader sequences (42.4% vs. 59.5%, respectively), which may be related to specific mechanisms of expression modulation (Pesole et al., 1997, 2000; Yángüez et al., 2013). Transcripts with high GC content or long CDSs have been shown to be poorly translated in drought stress in samples derived from *Arabidopsis* and maize plants (Kawaguchi and Bailey-Serres, 2005; Lei et al., 2015). Similarly, high 5′ UTR GC content in correlates inversely with polysome loading during hypoxia stress, ozone-induced oxidative stress and sudden darkness (Branco-Price et al., 2005; Juntawong and Bailey-Serres, 2012; Puckette et al., 2012). This may partially explain why transcripts encoding ribosomal proteins (which generally have high GC density in their 5′ UTRs) were found to be translationally downregulated in these various stress experiments (Muench et al., 2012). The 5′ UTRs of heat shock proteins *Arabidopsis* and maize have been

shown to be essential for promoting efficient translation during heat shock (Dinkova et al., 2005; Matsuura et al., 2008). In cold stress, data suggest that a different mechanism is at play -

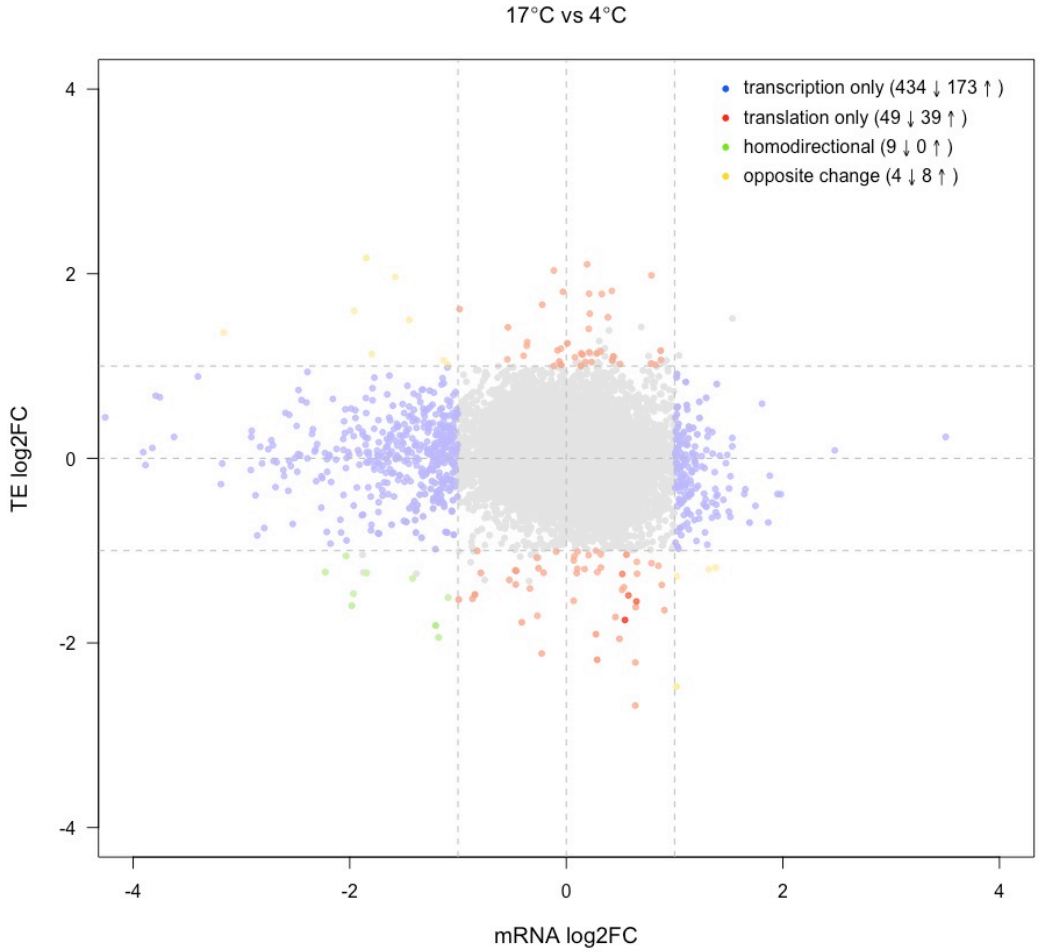


Figure 6.1 Differentially expressed genes after cold stress. RiboSeq and RNASeq libraries from 4 °C treated *Arabidopsis thaliana* plants were compared with libraries from 17 °C treated plants using Xtail. Grey lines indicate a log2 fold change greater than or equal to 1 (or less than -1). Coloured dots indicate a given gene had a FDR less than the threshold of 0.05. The grey dots are genes for which the estimated FDR value did not pass the threshold.

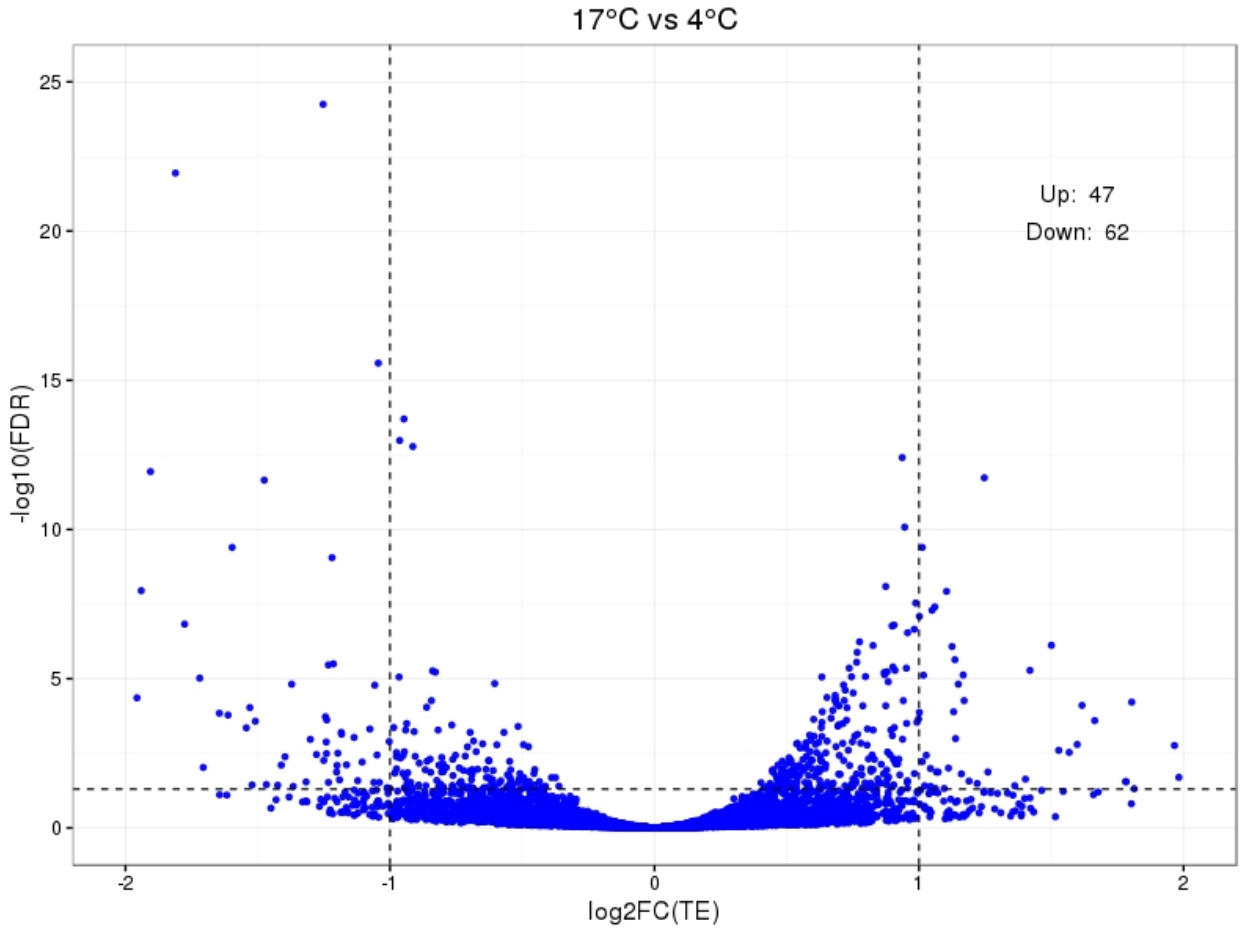


Figure 6.2 Volcano plot of differentially expressed genes after cold shock. Read counts from 4 °C and 17 °C libraries were compared using Xtail and filtered according to a log2 fold change in translation efficiency greater than 1 (or less than -1) and a FDR less than 0.05. The horizontal dashed line indicates the FDR cut-off of 0.05 on a –log10 scale. The vertical dash lines indicate the log2 fold change cut-offs for translation efficiency.

pathway description	observed gene count	false discovery rate
translation	9	5.71E-05
organonitrogen compound biosynthetic process	11	0.000195

Table 6.1 GO categories enriched in differentially translated genes after cold shock. Differentially expressed genes that passed FDR and log2 fold change filtering criteria in the comparison between the 4 °C and 17 °C libraries were analysed using StringDB (Szklarczyk et al., 2015). A list of enriched GO terms was downloaded and is presented here.

pathway description	observed gene count	false discovery rate
photosynthesis	16	4.57E-10
oxidation-reduction process	24	0.00158
plastid organization	8	0.0029
protein complex biogenesis	8	0.0137
generation of precursor metabolites and energy	9	0.0143
amide biosynthetic process	12	0.0143
organonitrogen compound metabolic process	21	0.0143
translation	11	0.0169
organonitrogen compound biosynthetic process	17	0.0175
electron transport chain	5	0.0382
response to karrikin	5	0.0433

Table 6.2 GO categories enriched at the transcript level after cold shock. Genes that were differentially expressed at the level of transcript abundance in the 4 °C and 17 °C libraries were analysed using StringDB.

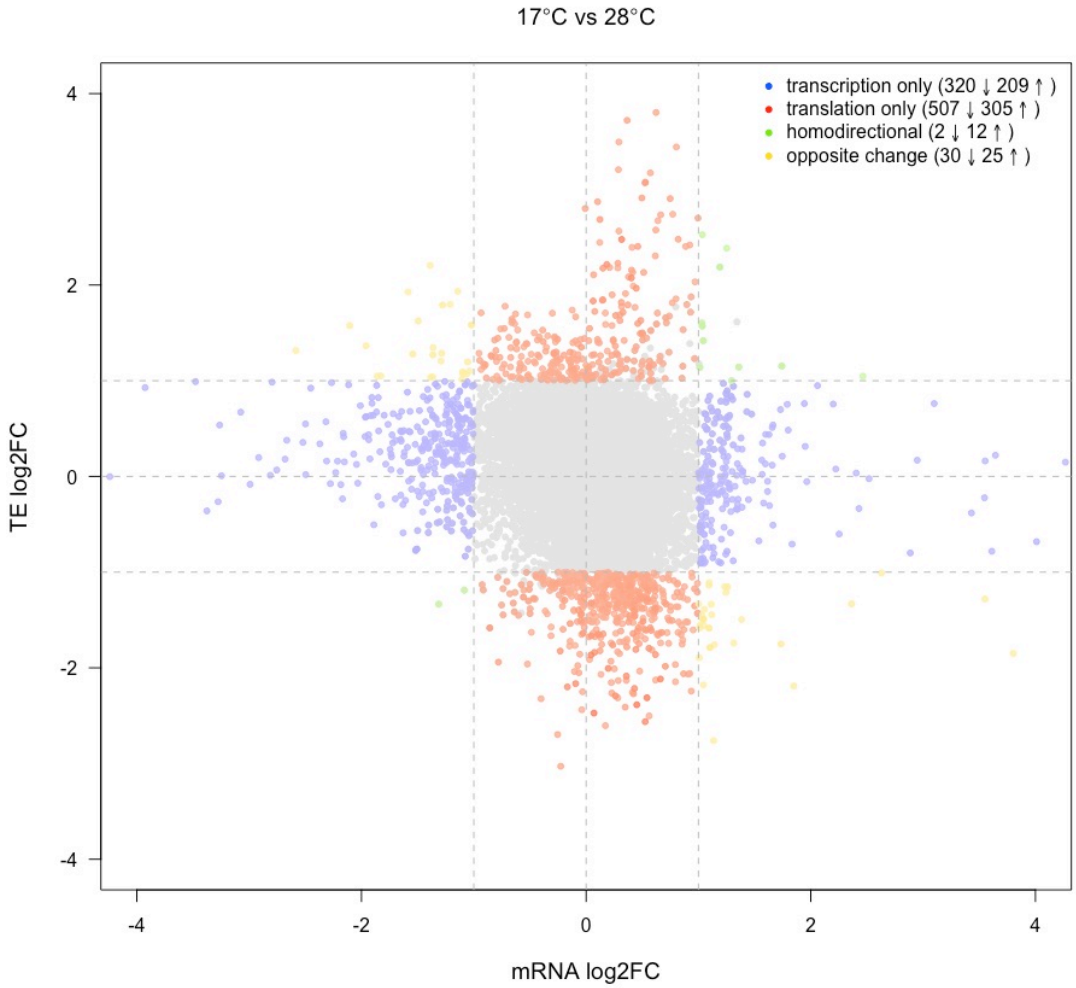


Figure 6.3 Differentially expressed genes after heat shock. RiboSeq and RNASeq libraries from 28 °C treated *Arabidopsis thaliana* plants were compared with libraries from 17 °C plants using Xtail. Grey lines indicates a log2 fold change greater than or equal to 1 (or less than -1). Grey dots indicate a given gene had a FDR greater than the threshold of 0.05.

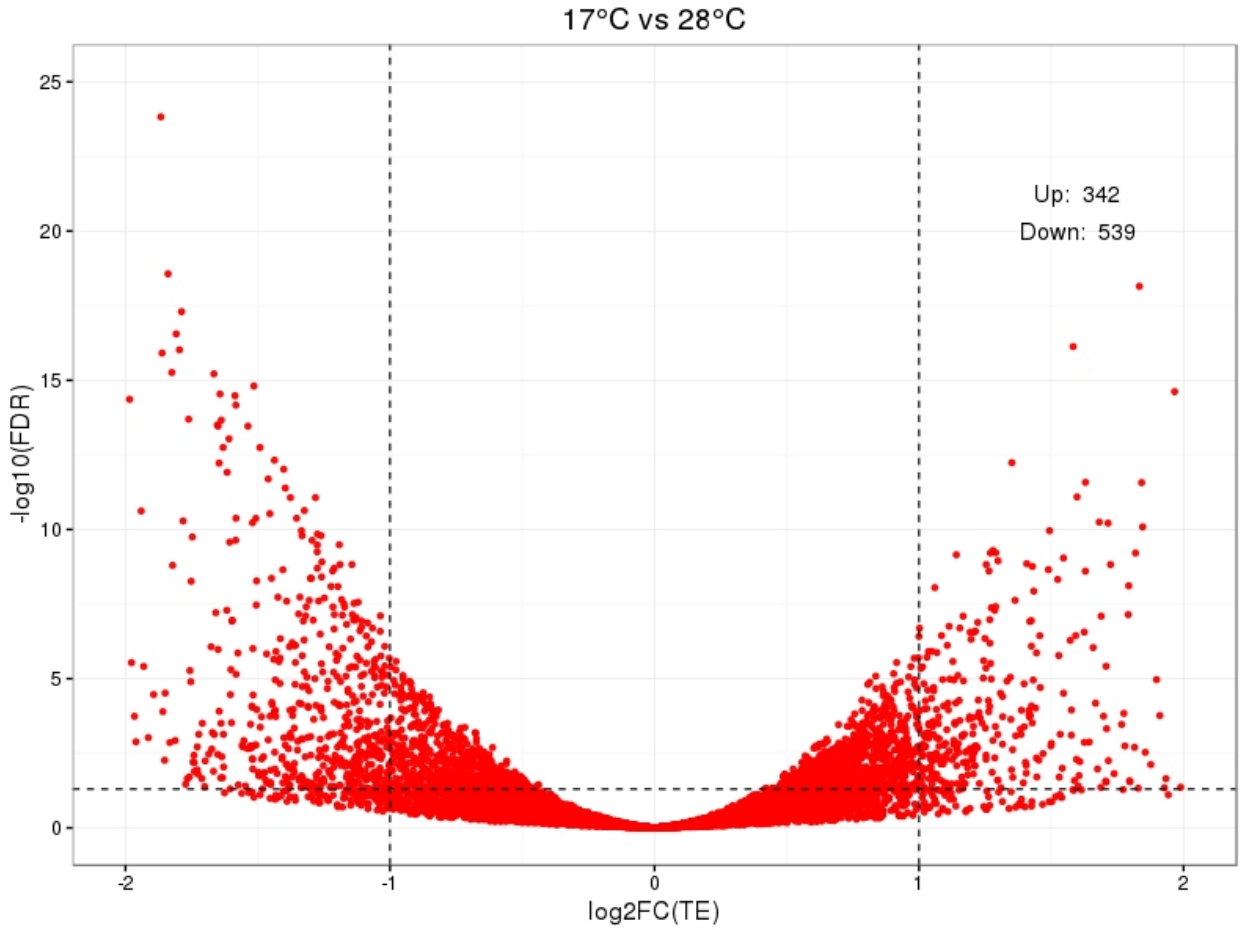


Figure 6.4 Volcano plot of differentially expressed genes after heat stress. Read counts from 28 °C and 17 °C libraries were compared using Xtail, and filtered according to a log2 fold change in translation efficiency greater than 1 (or less than -1) and a FDR less than 0.05. The horizontal dashed line indicates the FDR cut-off of 0.05 on a –log10 scale. The vertical dashed lines indicate the log2 fold change cut-offs for translation efficiency.

pathway description	observed gene count	false discovery rate
response to chitin	12	1.25E-06
gene expression	62	3.87E-06
response to oxygen-containing compound	34	8.51E-05
regulation of gene expression	50	0.000154
response to nitrogen compound	13	0.000212
translation	18	0.000239
regulation of transcription, DNA-templated	46	0.000278
response to wounding	10	0.000658
response to chemical	50	0.00072
response to organic substance	40	0.000829
response to acid chemical	26	0.000998
response to endogenous stimulus	37	0.00106
response to inorganic substance	22	0.00125
response to ethylene	12	0.00316
response to bacterium	13	0.00326
response to salicylic acid	8	0.00522
response to water deprivation	11	0.00542
response to stress	48	0.0111
RNA metabolic process	44	0.0146
response to oxidative stress	13	0.0177

Table 6.3 GO categories enriched in differentially translated genes after heat stress. Differentially expressed genes that passed FDR and log2 fold change TE filtering criteria in the 28 °C and 17 °C libraries were analysed using StringDB (Szklarczyk et al., 2015).

pathway description	observed gene count	false discovery rate
response to stimulus	68	6.71E-05
response to stress	61	4.89E-11
gene expression	47	0.000114
cellular protein metabolic process	43	0.0012
RNA metabolic process	37	0.0107
regulation of transcription, DNA-templated	34	0.00637
regulation of RNA metabolic process	33	0.0138
response to temperature stimulus	30	1.39E-14
response to inorganic substance	26	2.09E-07
response to oxygen-containing compound	26	0.00102
response to heat	19	1.86E-12
response to oxidative stress	17	9.26E-06
response to light stimulus	16	0.000661
response to cold	12	0.0015
response to osmotic stress	12	0.0377
response to reactive oxygen species	11	6.32E-06
response to metal ion	11	0.0287
response to hydrogen peroxide	10	2.93E-08
response to nitrogen compound	8	0.0396
innate immune response	8	0.0466
response to chitin	6	0.0231
defense response, incompatible interaction	6	0.0459
heat acclimation	4	0.0195

Table 6.4 GO categories enriched at the transcript level after heat stress. Genes that were differentially expressed at the level of transcript abundance in the 28 °C and 17 °C libraries were analysed using StringDB.

high GC content in the 5′ UTR was prevalent in CSP1-associated mRNAs, including biogenesis factors (Juntawong et al., 2013). Motivated by these meta-analyses, we sought to survey our data to see if we could see any distinctive correlations between the aforementioned transcript features and changes in translation efficiency or transcript abundance during heat or cold shock. We plotted the normalized fold-change values from our differential expression analyses against the GC content of 5′ UTRs, 5′ UTR length, 3′ UTR length and (annotated) CDS length, excluding any transcripts for which the UTR/CDS region was labelled as having a length of 0 nt or for which there were fewer than 50 RPFs mapping in all of the temperature conditions (
Figure 6.5,

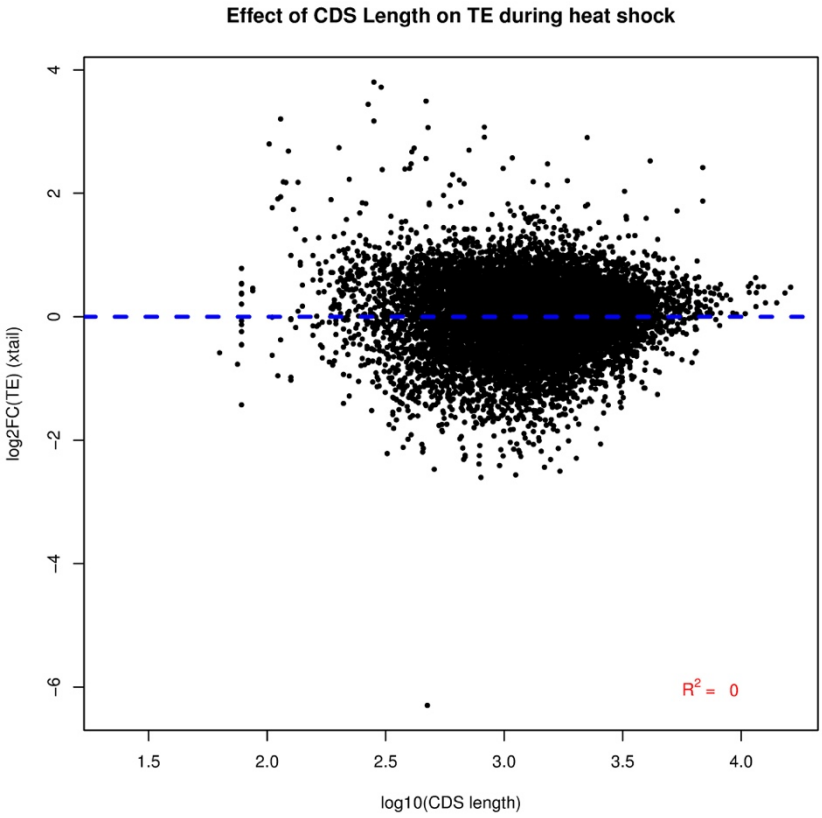
Figure 6.6,

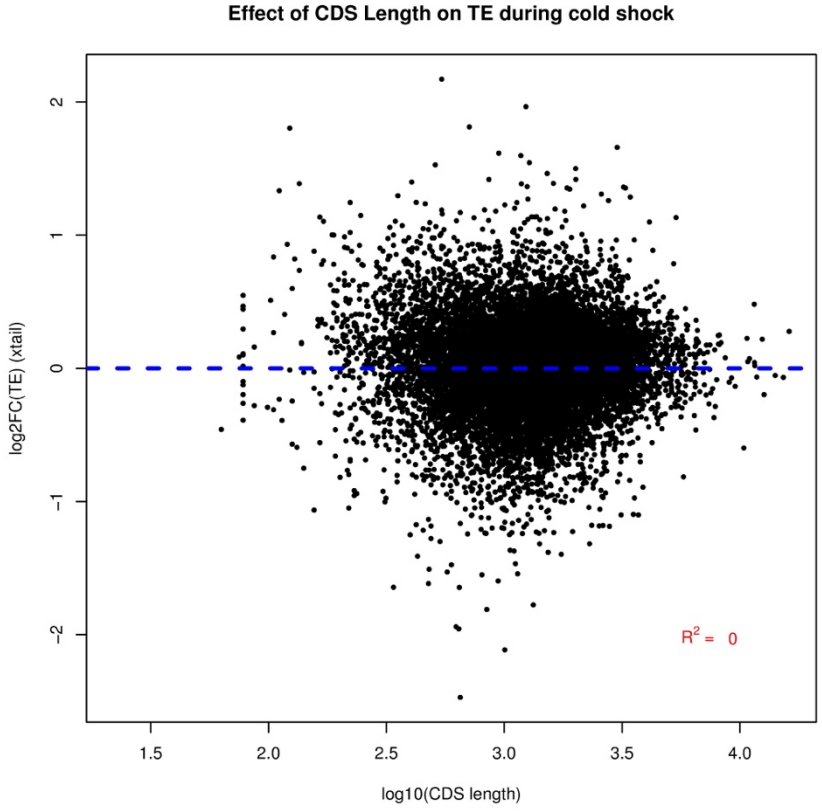
Figure 6.7). These analyses did not yield any discernible clustering or strong correlation coefficients ($R^2 \leq 0.03$) that would suggest sensitivity or recalcitrance to heat shock or cold shock.

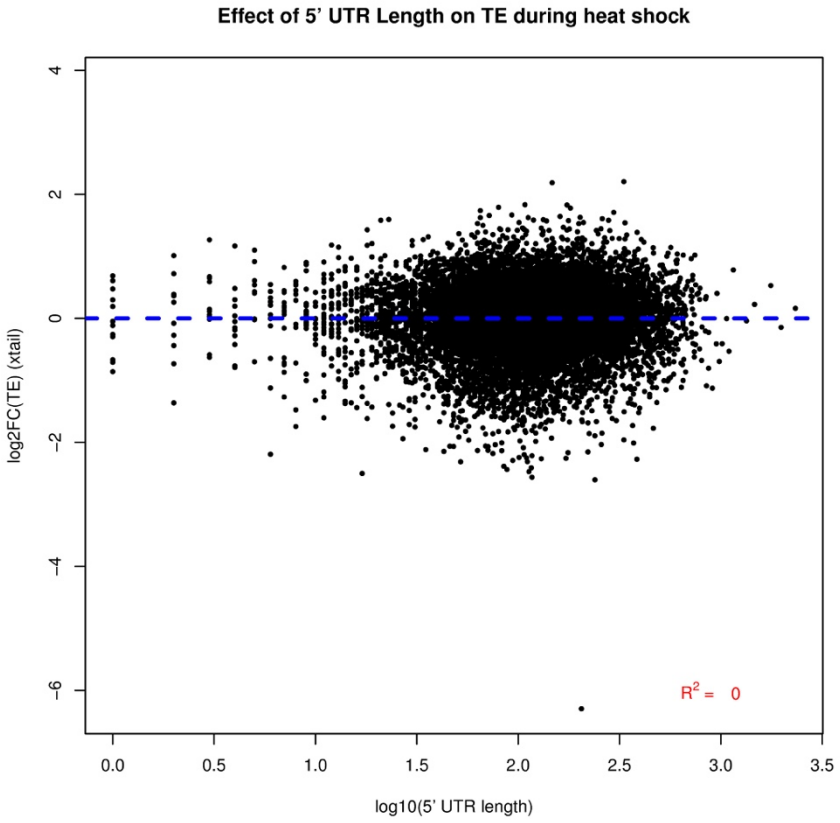
6.2.3 Effect of codon usage on gene expression

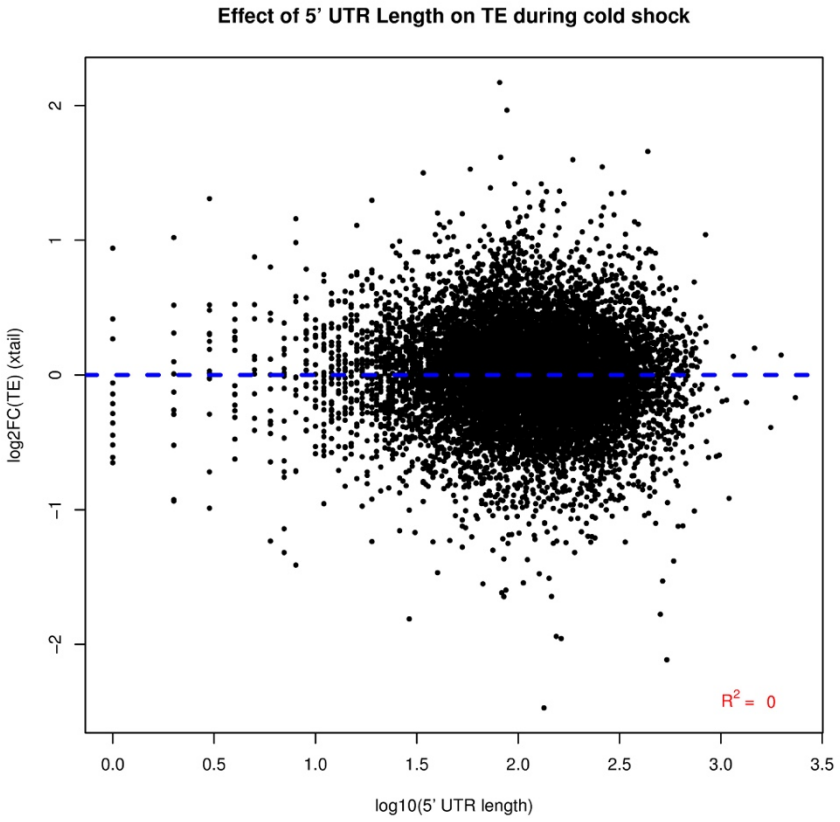
Synonymous codons are used with unequal frequencies in *Arabidopsis thaliana* and many other genomes (Lee et al., 2010). Numerous possible causes have been proposed for this phenomenon, including selective pressures to optimise for translational and transcriptional efficiency due to rapid organismal growth rates (Lee et al., 2010; Sharp et al., 2005), different rates of DNA mutation in the leading and lagging strands during replication (McLean et al., 1998) and the hydrophobicity values of proteins (D'Onofrio et al., 1999). Previous meta-analyses reported that a relative abundance of rare codons at the beginning of coding sequences can decrease initial translation rates and reduce bio-energetic demands (Tuller et al., 2010a). The codon adaption index (CAI) is a widely used metric of codon usage bias which measures usage of frequent codons (CAI = 1) or uncommon codons (CAI = 0) within a gene (Sharp and Li, 1987). Ribosome proteins, transcription factors and translation factors, as well as other highly expressed genes, frequently have high CAI values (Lee et al., 2010). We undertook a meta-analysis of transcript codon bias to see if there were any groups of genes that were translated with high efficiency during heat or cold stress and were enriched for either rare or abundant codons. *Arabidopsis* specific codon usage values for nuclear, mitochondrial and

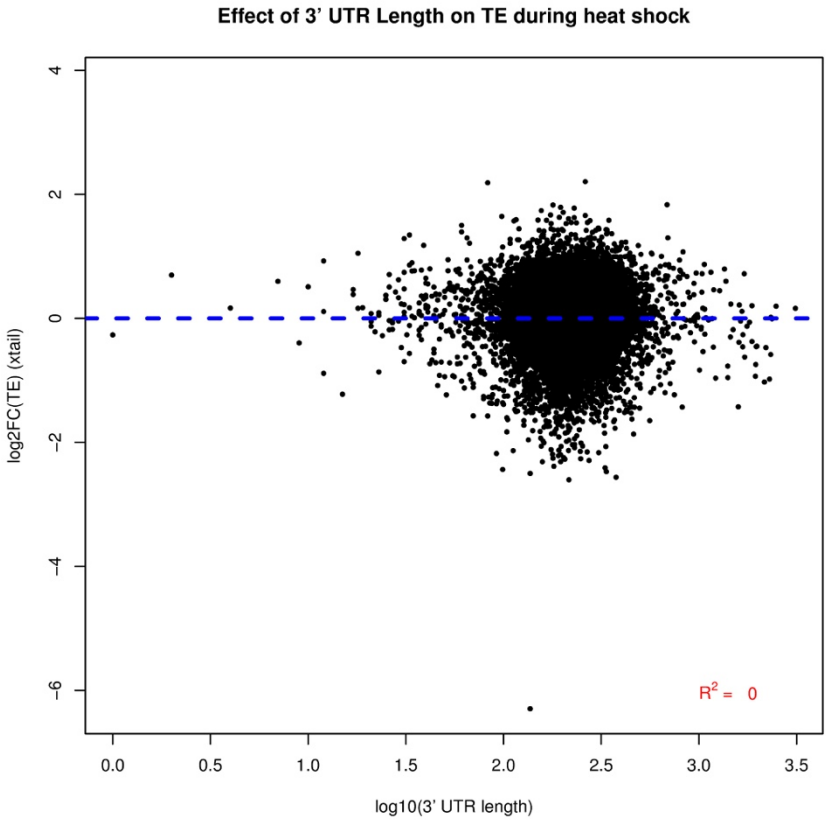
chloroplast genes were obtained from the Codon Usage Database, and were used to calculate CAI values for each CDS (Nakamura et al., 2000). CAI values were compared against changes in translational efficiency during heat shock and cold shock for both nucleus and chloroplast encoded transcripts (Figure 6.8). No clear separation or clustering was visible in the comparison. We also performed hierarchical clustering of individual codons in translationally


Effect of CDS Length on TE during heat shock
log2FC(TE) (xtail)
4
2
0
−2
−4
−6
R2 = 0
1.5
2.0
2.5
3.0
3.5
4.0
log10(CDS length)


Effect of CDS Length on TE during cold shock
log2FC(TE) (xtail)
2
1
0
−1
−2
R2 = 0
1.5
2.0
2.5
3.0
3.5
4.0
log10(CDS length)


Effect of 5' UTR Length on TE during heat shock
log2FC(TE) (xtail)
log10(5' UTR length)
R2 = 0


Effect of 5' UTR Length on TE during cold shock
log2FC(TE) (xtail)
log10(5' UTR length)
R2 = 0


Effect of 3' UTR Length on TE during heat shock
log2FC(TE) (xtail)
log10(3' UTR length)
R2 = 0

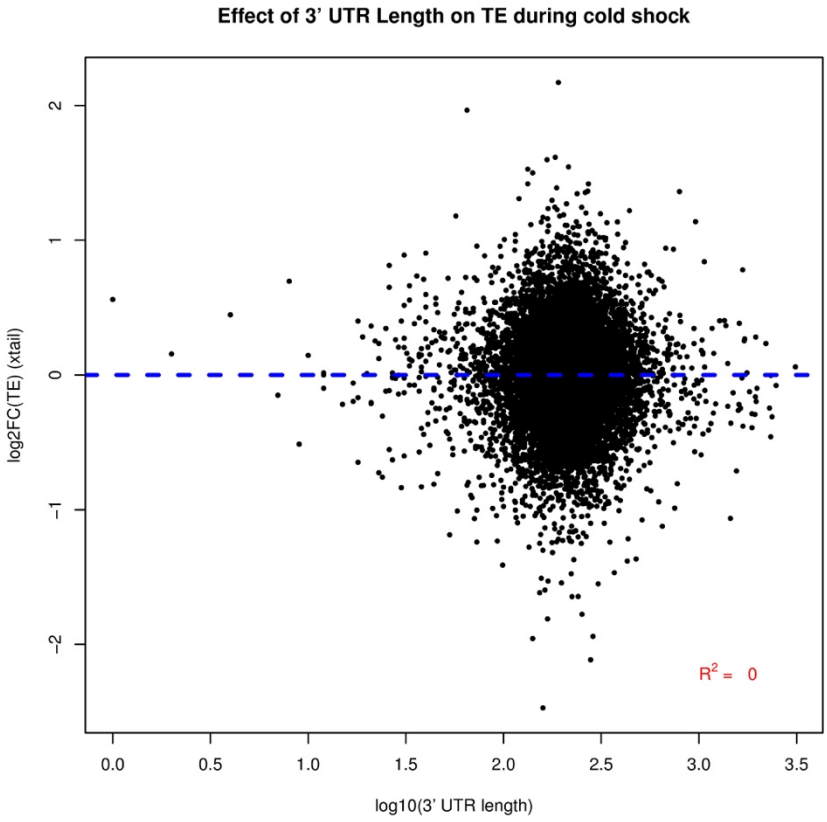
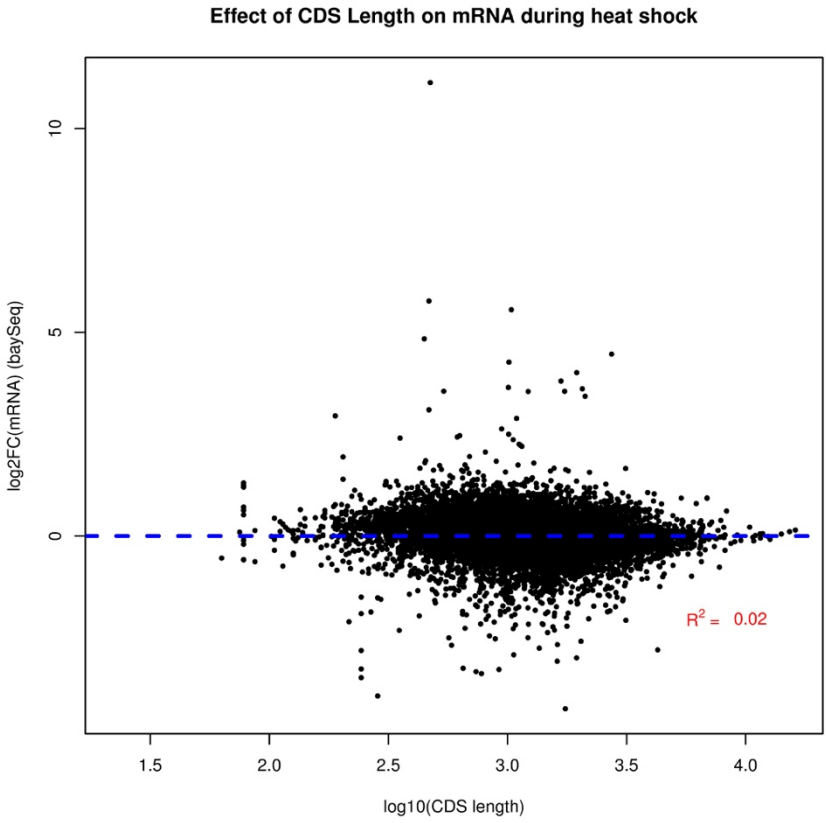
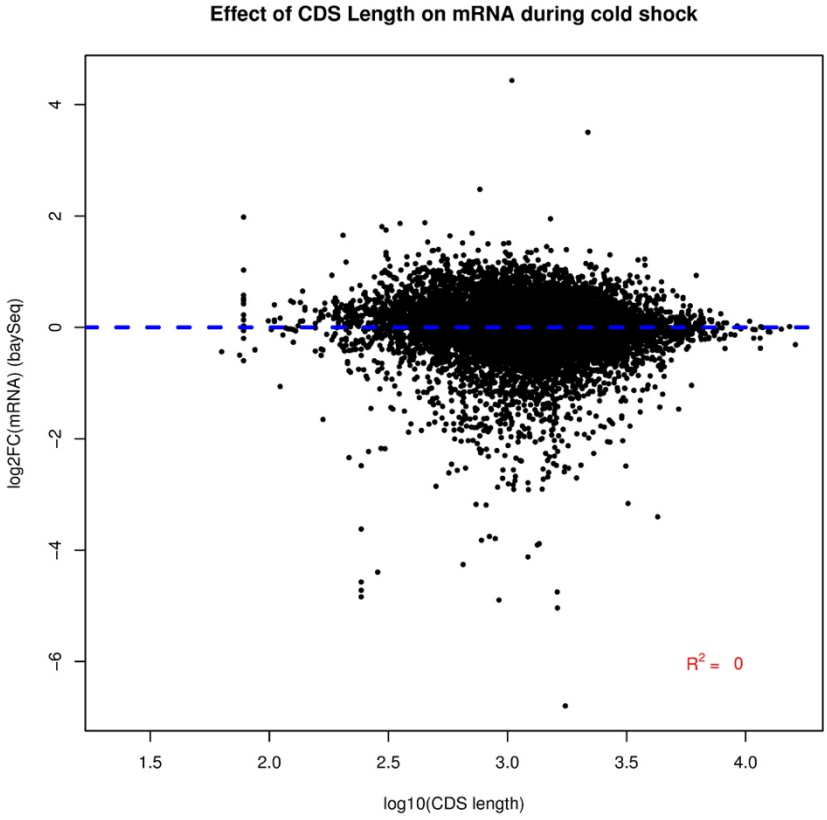
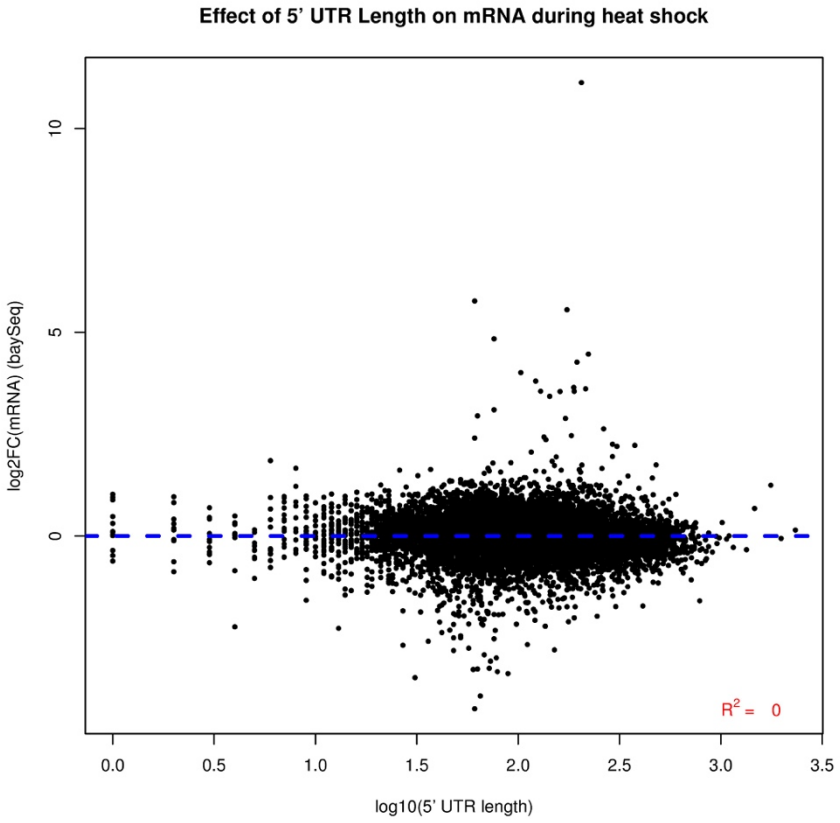
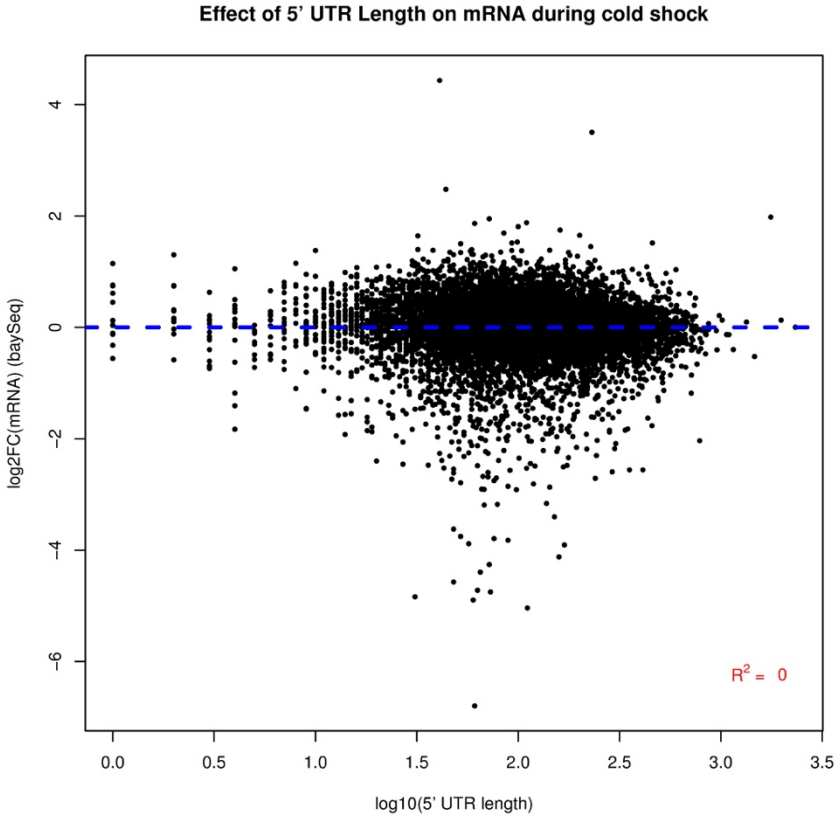

Effect of 3' UTR Length on TE during cold shock
log2FC(TE) (xtail)
log10(3' UTR length)
R2 = 0

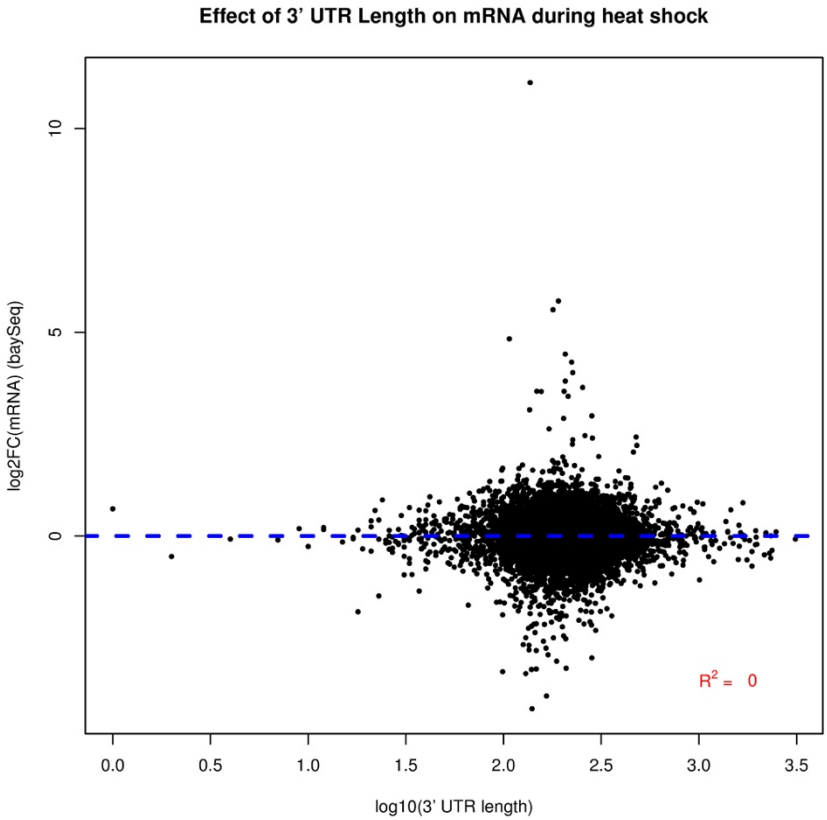
Figure 6.5 Comparison of transcript feature length and change in TE on temperature stress. Comparison of CDS length, 5′ UTR length and 3′ UTR length with changes in translation efficiency during heat stress (17 °C vs 28 °C) or cold shock (17 °C vs 4 °C). Transcript features are on a log10 scale. Before commencing these analyses, genes with fewer than 50 read counts in all temperature conditions were removed. Reads were mapped to the TAIR10 *Arabidopsis thaliana* transcriptome using Bowtie.


Effect of CDS Length on mRNA during heat shock
log2FC(mRNA) (baySeq)
10
5
0
1.5
2.0
2.5
3.0
3.5
4.0
log10(CDS length)
R2 = 0.02


Effect of CDS Length on mRNA during cold shock
log2FC(mRNA) (baySeq)
4
2
0
−2
−4
−6
1.5
2.0
2.5
3.0
3.5
4.0
log10(CDS length)
R2 = 0


Effect of 5' UTR Length on mRNA during heat shock
log2FC(mRNA) (baySeq)
10
5
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
log10(5' UTR length)
R2 = 0


Effect of 5' UTR Length on mRNA during cold shock
log2FC(mRNA) (baySeq)
4
2
0
−2
−4
−6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
log10(5' UTR length)
R2 = 0


Effect of 3' UTR Length on mRNA during heat shock
log2FC(mRNA) (baySeq)
10
5
0
R2 = 0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
log10(3' UTR length)

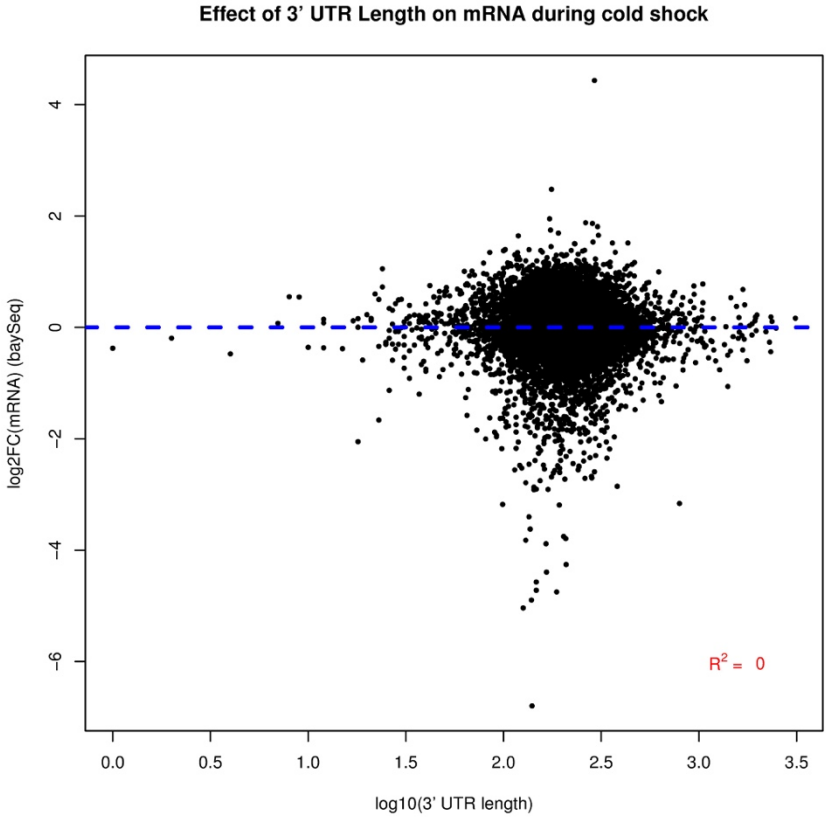
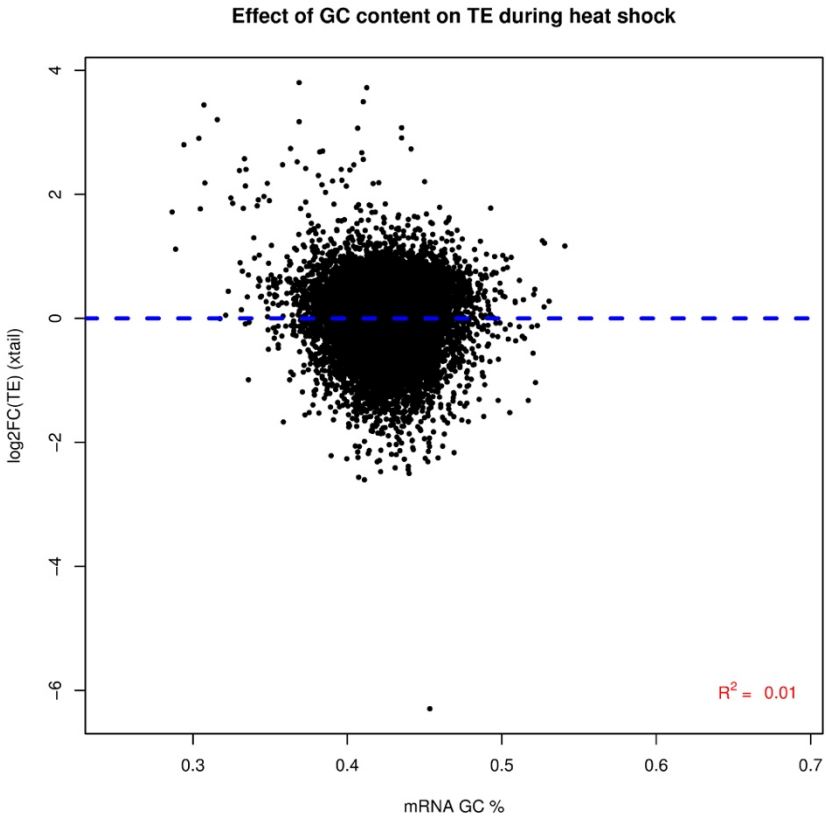
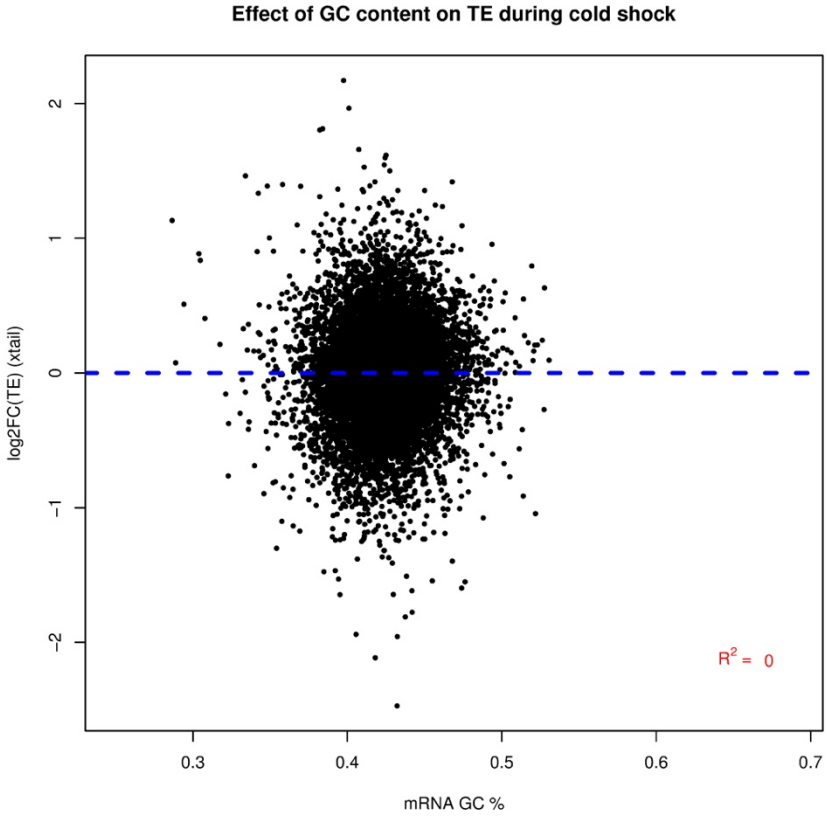
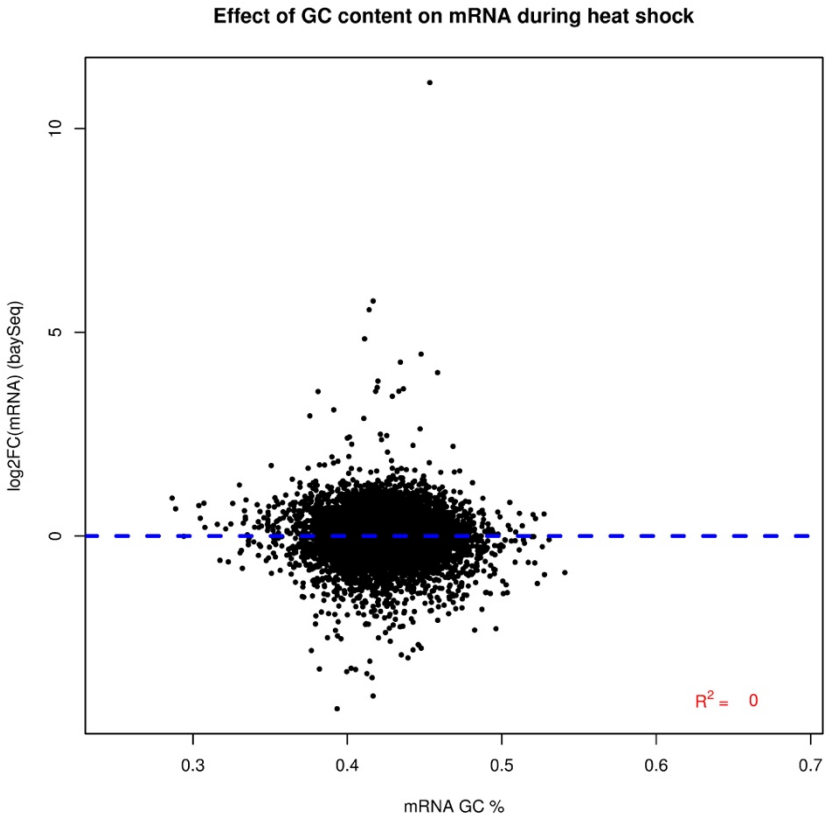

Effect of 3' UTR Length on mRNA during cold shock
log2FC(mRNA) (baySeq)
4
2
0
−2
−4
−6
R2 = 0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
log10(3' UTR length)

Figure 6.6 Comparison of transcript feature length and change in transcript abundance. Comparison of CDS length, 5′ UTR length and 3′ UTR length with changes in transcript abundance during heat stress (17 °C vs 28 °C) or cold shock (17 °C vs 4 °C). Transcript feature lengths were plotted on a log10 scale. Before commencing theses analyses, genes with fewer than 50 read counts in all temperature conditions were removed.


Effect of GC content on TE during heat shock
log2FC(TE) (xtail)
mRNA GC %
R2 = 0.01
4
2
0
−2
−4
−6
0.3
0.4
0.5
0.6
0.7


Effect of GC content on TE during cold shock
log2FC(TE) (xtail)
mRNA GC %
R2 = 0
2
1
0
−1
−2
0.3
0.4
0.5
0.6
0.7


Effect of GC content on mRNA during heat shock
log2FC(mRNA) (baySeq)
10
5
0
0.3
0.4
0.5
0.6
0.7
mRNA GC %
R2 = 0

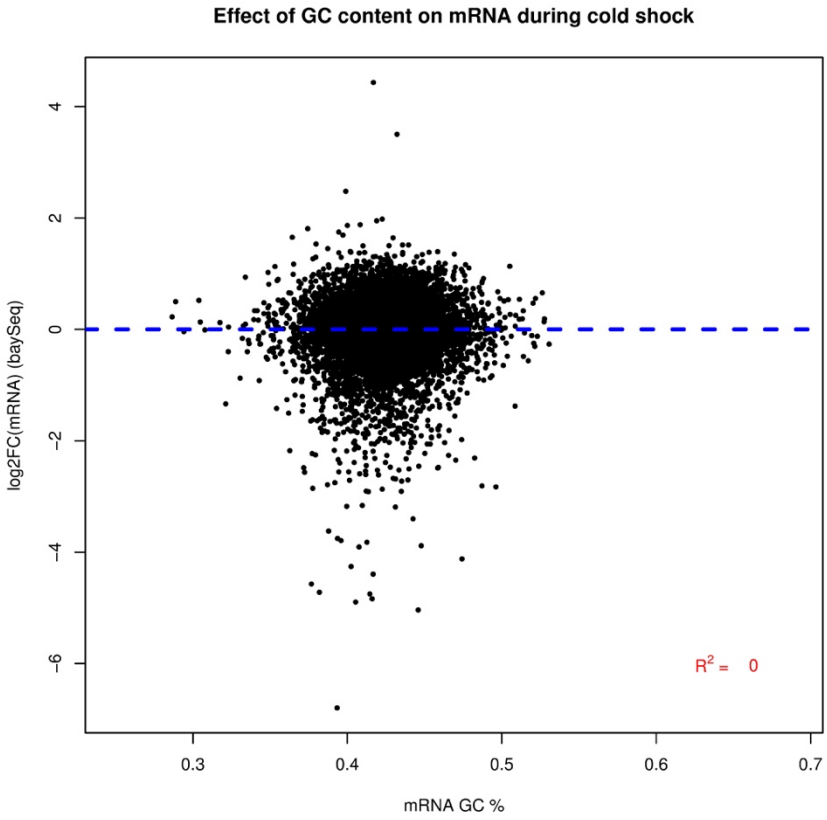
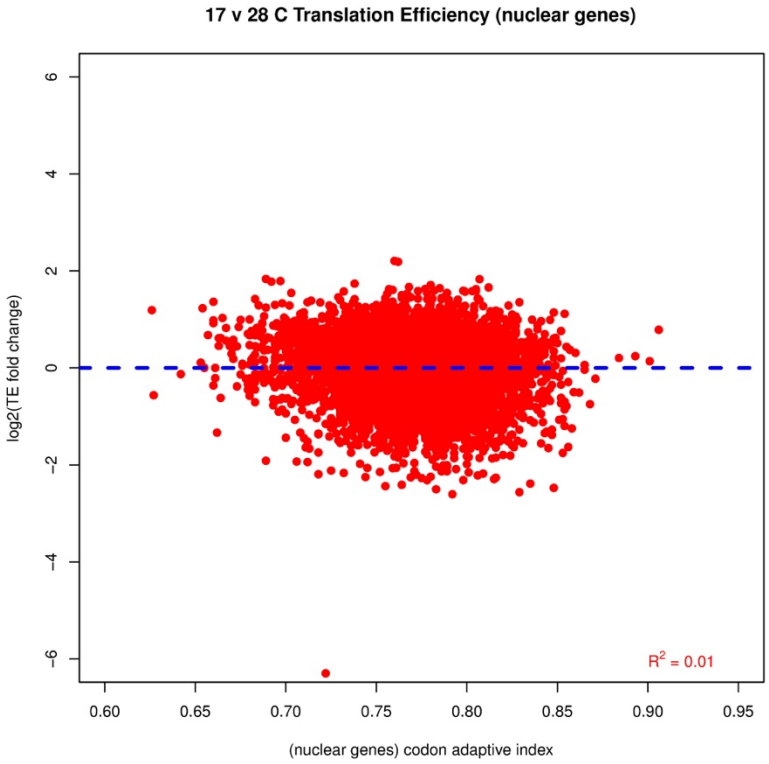
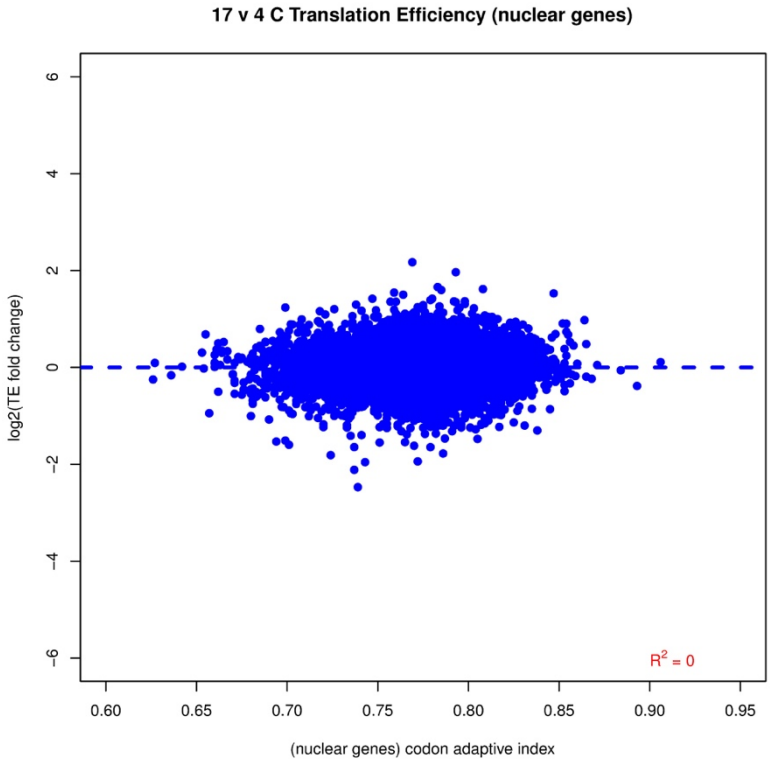
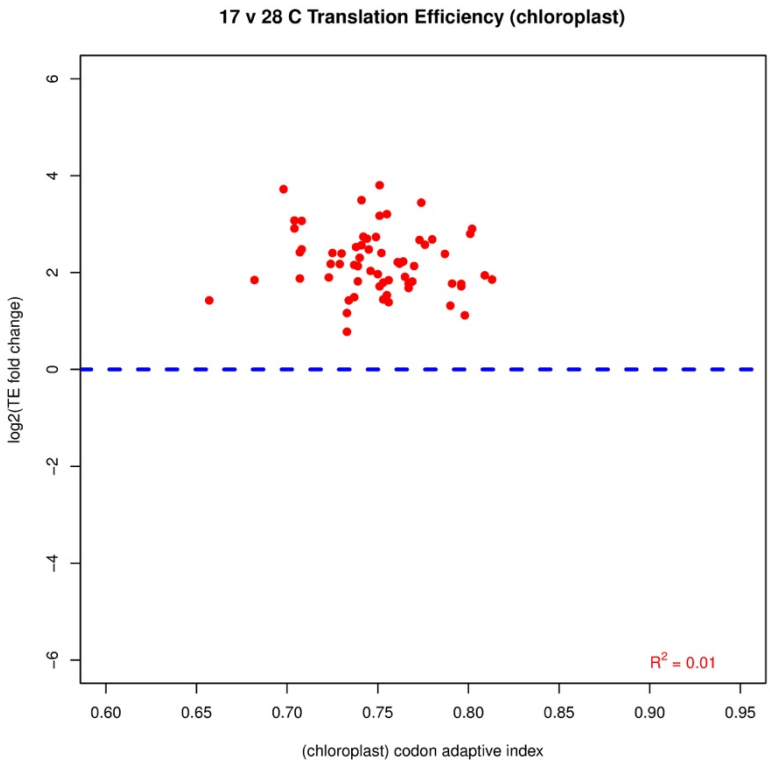

Effect of GC content on mRNA during cold shock
log2FC(mRNA) (baySeq)
4
2
0
−2
−4
−6
0.3
0.4
0.5
0.6
0.7
mRNA GC %
R2 = 0

Figure 6.7 Comparison of transcript GC content and changes in TE or mRNA abundance. Comparison of transcript GC content (including UTRs) with changes in translation efficiency or transcript abundance during heat stress (17 °C vs 28 °C) or cold shock (17 °C vs 4 °C). Before commencing theses analyses, genes with fewer than 50 read counts in all temperature conditions were removed.


17 v 28 C Translation Efficiency (nuclear genes)
log2(TE fold change)
6
4
2
0
−2
−4
−6
R2 = 0.01
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
(nuclear genes) codon adaptive index


17 v 4 C Translation Efficiency (nuclear genes)
log2(TE fold change)
6
4
2
0
−2
−4
−6
R2 = 0
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
(nuclear genes) codon adaptive index



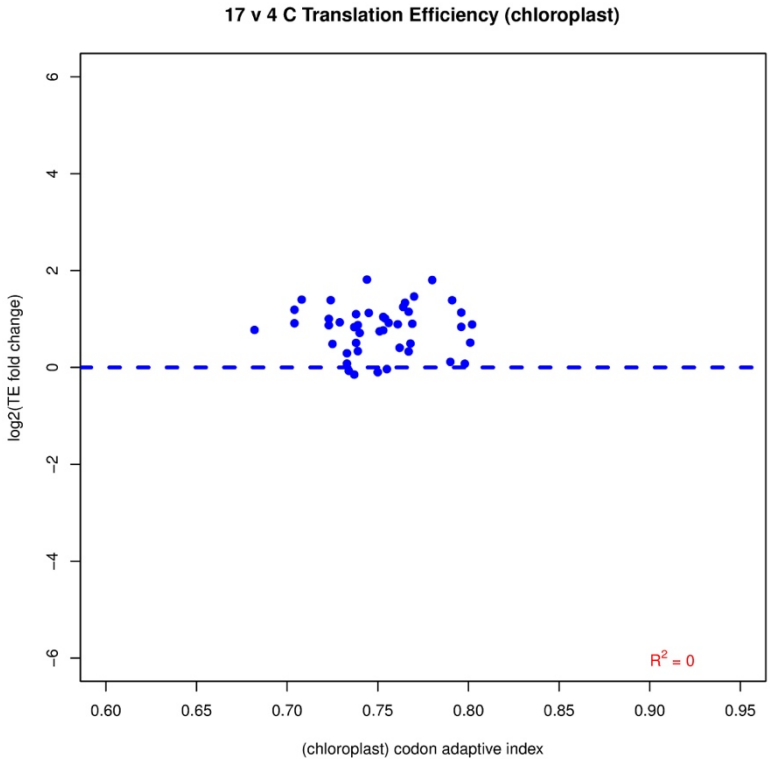


Figure 6.8 Comparison of CAI values and changes in TE. Comparison of CAI values with changes in translation efficiency during heat stress (17 °C vs 28 °C) or cold shock (17 °C vs 4 °C) in nuclear gene encoded transcripts (first two panels) or chloroplast encoded transcripts (last two panels). Before commencing theses analyses, genes with fewer than 50 read counts in all temperature conditions were removed.

modulated genes (Figure 6.9,
Figure 6.10). While certain codons were frequently used in these ORFs (such as the lysine AAG codon), we saw both rare and abundant codons enriched in both datasets, and we could not see any discernible difference in clustering between translationally upregulated or downregulated genes at 4 °C or 28 °C.

6.3 Identification of uORFs in *Arabidopsis*

As alluded to in Chapter 1, *Arabidopsis thaliana* and many other eukaryotic genomes have been found to contain numerous upstream ORFs that may encode micro-peptides and/or regulate translation of downstream ORFs (Tran et al., 2008). ~20% of *Arabidopsis thaliana* transcripts contain uORFs, including 30 gene families which contain uORFs that have been shown to encode evolutionarily conserved micro-peptides (Hayden and Jorgensen, 2007; Jorgensen and Dorantes-Acosta, 2012; Vaughn et al., 2012). Many of these uORFs have been shown to function in response to specific environmental stimuli (Hanfrey et al., 2002; Wiese, 2004; Nishimura et al., 2005; Imai et al., 2006; Alatorre-Cobos et al., 2012; Rosado et al., 2012). The AtbZIP11 gene, which regulates amino acid and sugar metabolism, encodes a mRNA that contains four uORFs (Wiese, 2004; Hanson et al., 2008; Thalor et al., 2012). The translation of the main ORF is repressed by sucrose through a peptide that is encoded by the second uORF (Rahmani et al., 2009). Transcripts that contain translated uORFs have been shown to be translated less efficiently during darkness, relative to transcripts with inactive uORFs (Liu et al., 2013b). Ribosome profiling of maize seedlings during drought stress indicated potential translation of 3063 uORFs from 2558 distinct mRNAs, some of which may also function as a stress sensing and adaptation mechanism during drought stress indicated potential translation of 3063 uORFs from 2558 distinct mRNAs, some of which may also function as a stress sensing and adaptation mechanism (Lei et al., 2015). Previous studies have shown that plant uORFs can regulate translation of specific transcripts in response to small molecules such as polyamines (Hanfrey et al., 2005). We decided to survey our RiboSeq data to determine if we could detect evidence of differential uORF translation in response to heat shock or cold stress, as well as to analyse start codon usage for particular uORFs.

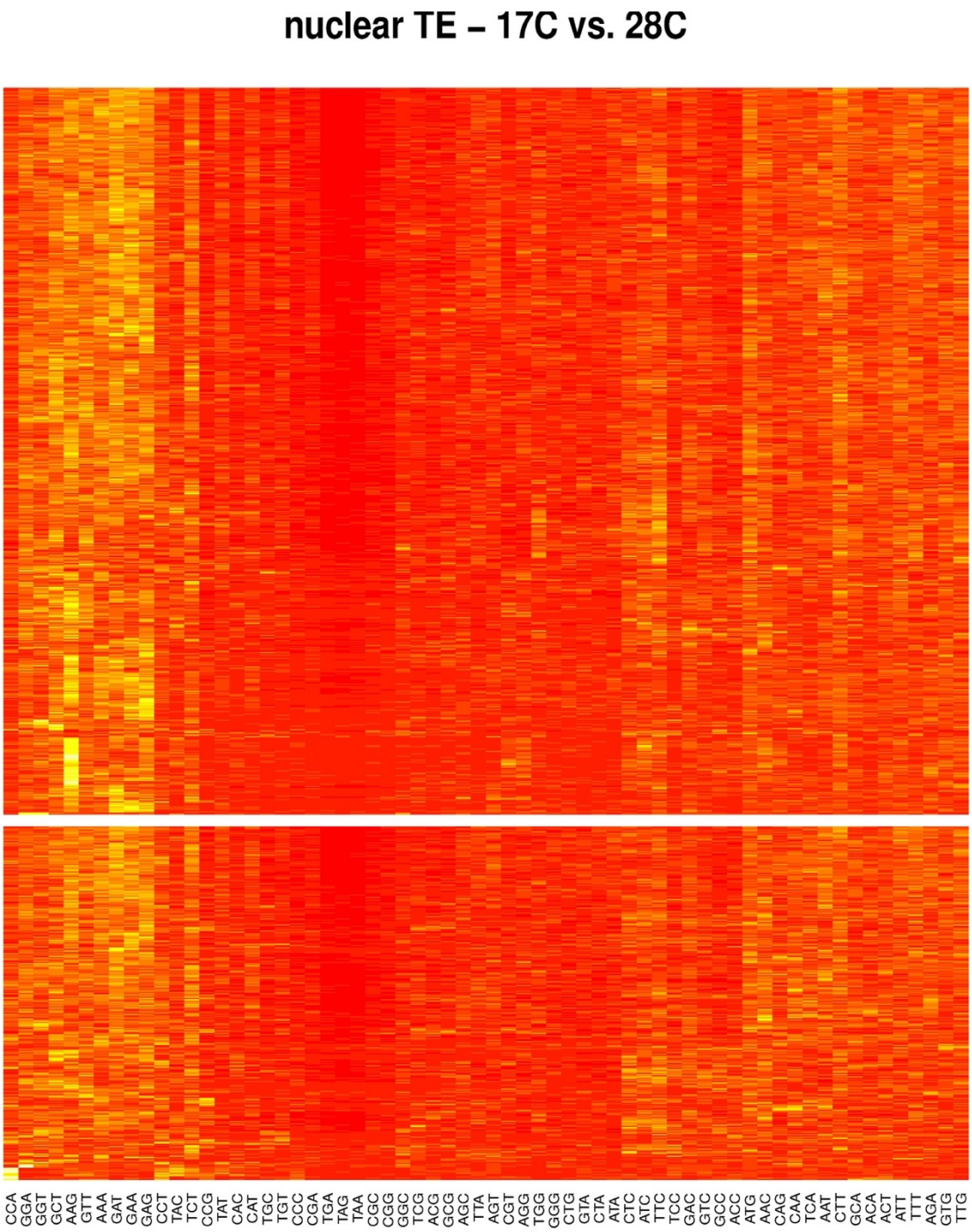


Figure 6.9 Heatmap of codon utilization in differentially translated mRNAs during heat shock. Hierarchical clustering of codon frequency in the coding sequences of translationally upregulated (top panel) or downregulated (bottom panel) nuclear gene encoded transcripts during heat shock (28 °C).

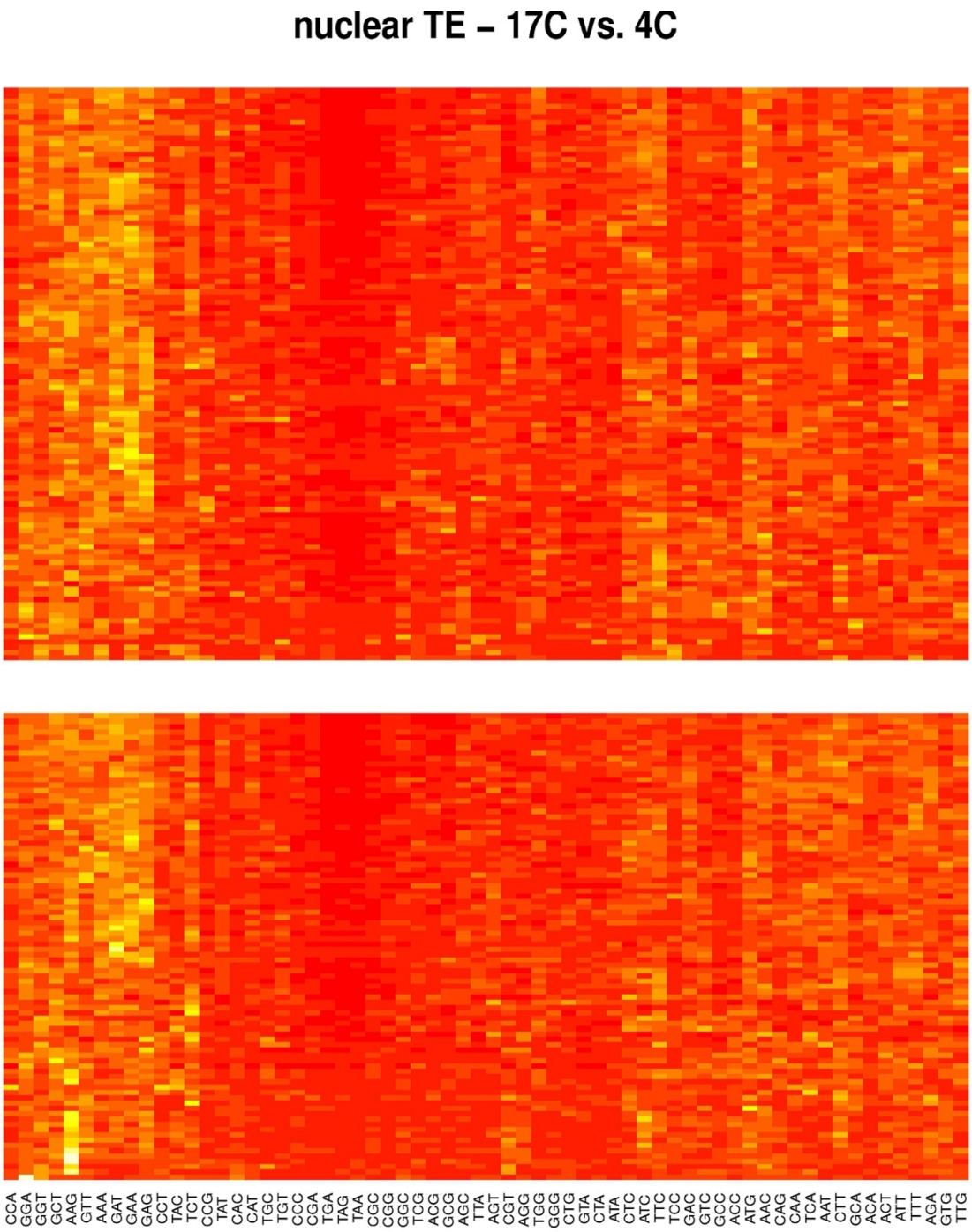


Figure 6.10 Heatmap of codon usage in differentially translated mRNAs during cold shock. Hierarchical clustering of codon frequency in the coding sequences of translationally upregulated (top panel) or downregulated (bottom panel) genes during cold shock (4 °C).

6.3.1 Additions to RiboSeqR package

RiboSeqR is a publicly available R package that is designed to detect translated ORFs using framing and unique read counts from ribosome profiling data (Chung et al., 2015). We decided to extend the functionality of this package by incorporating annotation information that would allow for the identification of translated sequences that occur upstream of main ORFs.

A new function titled '*finduORF*' was introduced which first finds all putative ORFs and then filters for those that have a start codon that occurs upstream of the main ORF start codon on a given transcript. The definition of main ORFs is user-defined through a modified *GRanges* object, and can use any reference annotation or a custom genome annotation. After a list of putative uORFs has been generated, the algorithm checks for framing and filters the sequences by number of unique read counts (to avoid characterisation of a single, highly abundant RPF peak as a potential uORF). The function then returns a *GRanges* object that lists the start codon, stop codon and sequence coordinates of all potential uORFs that pass the filtering criteria. The function can be applied recursively in order to find 5′ leaders that contain multiple successive uORFs. The algorithm was validated on the public data set available with the RiboSeqR repository.

6.3.2 Identification of AUG and non-AUG initiated short ORFs in 5′ leader sequences

We used the *finduORF* algorithm with the TAIR10 *Arabidopsis thaliana* annotation on each of our temperature RiboSeq libraries. Given previous analyses of *Arabidopsis* RPFs with RNase I (Chung et al., 2015), we defined 27-nt reads as being in the 0-frame, and 28-nt reads as being in the +1 frame. We ran the algorithm using eight different potential start codons (AUG, AUU, CUG, ACG, GUG, UUG, AUC, AUA). The minimum filtering thresholds were set to an average of 40 RPFs mapping to the given ORF and a minimum of 10 unique hits. The initial survey revealed a total of 13409 potential uORFs in the three RiboSeq libraries.

We tried to assess whether the presence of a predicted AUG initiated or non-AUG initiated uORF in a transcript contributed to selective modulation during heat shock or cold shock by plotting cumulative distribution functions (CDFs) of the calculated fold-change values and performing two-sided Kolmogorov-Smirnov test (Marsaglia et al., 2003; Figure 6.11). The estimated D-statistics did not indicate significant differences between the various distributions ($D \leq 0.05$; $p > 0.9$).

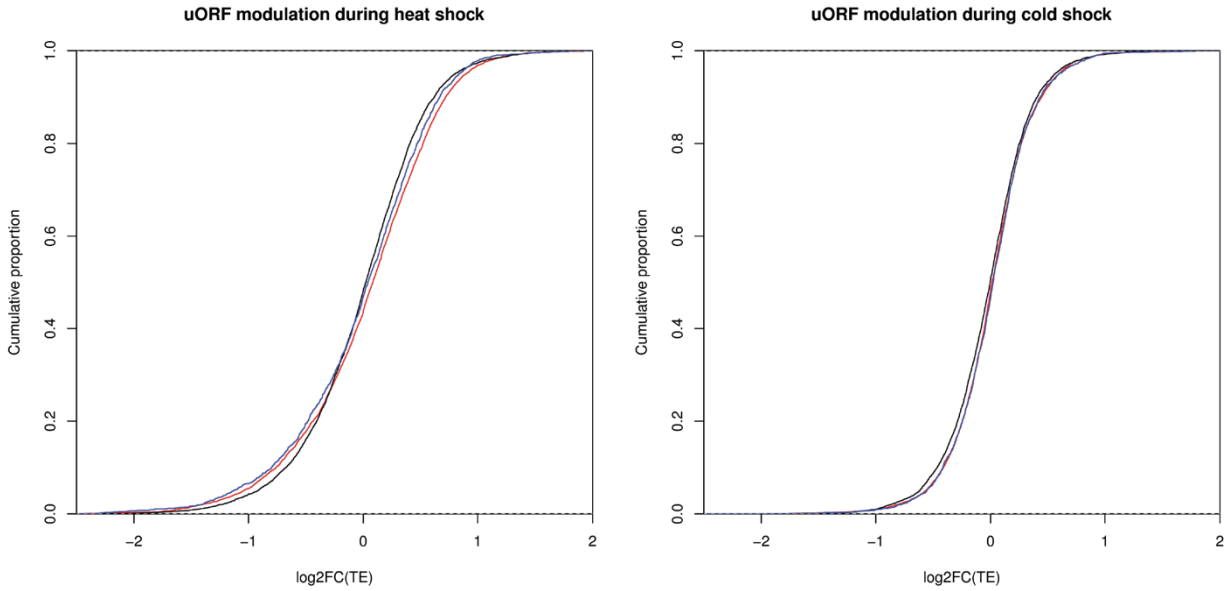


Figure 6.11 Cumulative density functions of differential translation in uORF-containing transcripts. The log2-fold changes in translational efficiency (as calculated by Xtail) for three groups of genes were plotted as cumulative distribution functions (CDFs). The red curves correspond to transcripts with predicted non-AUG uORFs, the black curves represent transcripts with predicted AUG-initiated uORFs and the blue curves represent genes with no predicted uORFs. A gene was included in the first two categories only if there was evidence of the uORF being translated based on the riboSeqR annotations. The left panel contains values generated by comparing the 17 and 28 °C samples; the right panel corresponds to fold-changes calculated by comparing the 17 and 4 °C samples. Two-sample Kolmogorov-Smirnov tests were used to compare the various CDFs; the estimated D-statistics did not suggest significant differences between the various distributions ($D \leq 0.05$; $p > 0.9$).

6.4 Discussion

This is the first high-resolution, genome-wide analysis of plant translation subject to temperature modulation. The differential expression analyses corroborate previous reports on the effects of heat shock and cold stress on gene expression in plants, as well as introducing several new observations that merit future investigation. Cold stress resulted in a translational upregulation of ribosome biogenesis factors and organo-nitrogen compound response genes, while poly(A) tail shortening genes were downregulated at both the transcript abundance and translational levels. Molecular imaging, tail sequencing and *in-situ* hybridization experiments will be needed to determine if cold-stress results in a transient reduction in transcript turnover as a means of preserving mRNA pools. Our data indicated an upregulation in heat shock protein and oxidative stress response factor expression, along with a reduction in photosynthetic and translation apparatus expression. These results are largely in line with previous analyses of heat shock effects on translation (Matsuura et al., 2010). Future work will include additional replicates for each temperature condition in order to improve the power of detection for the differential expression analyses. qRT-PCR and protein-based techniques will be used to validate the changes in genes of particular interest. Plants are unique in their selective use of an alternative eIF4 binding factor, eIFiso4F, to translate specific sets of mRNAs (Mayberry et al., 2011). This factor allows for more efficient translation initiation on transcripts with complex 5′ UTR secondary structure and with hypermethylated cap structures (Carberry et al., 1991). Recently developed methodologies allow for genome-wide analyses of RNA structure *in vivo*, and can be used to determine if the heat or cold-shock recalcitrant transcripts are partially regulated through the presence of unusual RNA structures in their leader sequences (Ding et al., 2014; Rouskin et al., 2014; Wan et al., 2014).

Meta-analyses of various transcript features such as codon adaptation indices, 5′ GC content, and 5′ UTR length did not show distinct clustering when compared with changes in translation efficiency during heat stress or cold stress. These analyses may have been affected by the high levels of uncertainty imposed by estimated changes in translation activity from single samples (with no replicates); future work will incorporate additional replicates into estimates of translation efficiency to see if any distinctive clustering is observable. It is possible also that at the specific temperature ranges used in this study (4, 17, 28 °C), the aforementioned intrinsic transcript features are not the primary means of selecting transcripts for continued translation, and that *trans*-acting factors, selective compartmentalization or other factors act as the primary modulators.

We extended existing software for annotation of translated sequences to survey uORF activity in the *Arabidopsis* transcriptome. A major limitation of our present dataset is the lack of initiation profiling (e.g. harringtonine or lactimidomycin) libraries, which makes it difficult to deconvolute different potential translation start sites that occur in the same reading frame. Future work will incorporate initiation profiling to test, and improve the resolution of uORF detection and analyse uORF usage in response to temperature modulation. In summary, this study highlights the importance of translation regulation in the context of plant temperature stress responses and introduces several observations that may be used to inform the development of heat or cold resistant crops.

Chapter Seven

Plenary Discussion

7 CHAPTER SEVEN: PLENARY DISCUSSION

7.1 Introduction

Ribosome profiling has been used to study gene expression and unusual regulatory mechanisms in a myriad of organisms. It offers an unprecedented level of breadth when surveying translation, though it is possible to over-interpret profiling data (Ingolia, 2016).

In this thesis, I describe several attempts to mine and extract useful information from several RiboSeq datasets. I customised and extended analyses for three viral systems and four cell lines, examined differential gene expression in MHV-infected murine cells and temperature-shocked *Arabidopsis* plants, and developed custom pipelines for dealing with several profiling specific challenges.

7.2 Genome-wide studies of viral and metazoan gene expression

Numerous previous studies measuring absolute levels of mRNA and proteins have revealed that dynamic changes in the transcriptome do not necessarily predict alterations in the proteome of cells in response to external cues (Piques et al., 2009; Vogel et al., 2010; Lee et al., 2011). This theme recurs throughout the analyses in Chapter 3 and the supplementary chapter.

In murine cells infected with MHV, the differential expression analyses and visualization of read alignments revealed genetic signatures consistent with the activation of the unfolded protein response, such as differential uORF utilization in Atf4 and increased translation of the main ORF. While there has been speculation in previous reports on the role of the Atf6 branch of the UPR upon MHV infection, our data suggest that this element is not upregulated translationally (Szegezdi et al., 2006; Bechill et al., 2008). A downstream effector of Atf6, the transcription factor Xbp-1, was not up-regulated transcriptionally but did exhibit a change in splicing consistent with production of the stress-activated variant of the protein.

Short-term temperature stress in *Arabidopsis* revealed scores of genes that were up- or downregulated at the level of translation efficiency but not transcript abundance (though many genes did exhibit a correlated response). Cold stress resulted in a translational upregulation of ribosome biogenesis factors and organo-nitrogen compound response genes, while poly(A) tail shortening genes were downregulated at both the transcript abundance and translational levels. Heat shock protein and oxidative stress response factor exhibited increased expression, along with a reduction in photosynthetic and translation apparatus expression. These results are largely in line with previous analyses of heat shock effects on translation (Matsuura et al.,

2010).

This thesis presents the first (to my knowledge) high-throughput analyses of DENV and HIV translation. Several large pause sites were detected, including one that occurs in the vicinity of an 11 codon stretch of the *tat* ORF that encodes a highly positively charged and proline-rich segment that may be involved in induced peptide-mediated stalling, and another immediately upstream of the predicted site of the HIV Rev response element secondary structure. A comparison of read densities in the *gag* and *pol* ORFs suggested that cellular ribosomes frameshift ~20% of the time at the *gag-pol* frameshift site, a higher estimate than those described by other methodologies. Ribosomes were also not seen to appreciably stall at the frameshift site, similarly to what was described for the MHV frameshift signal (Irigoyen et al., 2016). The various parts of the DENV polyprotein ORF were shown to be translated with roughly uniform efficiency, though major pauses were detected within NS2A and NS5-encoding segments.

7.3 Contributions to analysis of profiling data

During the course of the work on murine gene expression during coronavirus infection, a custom pipeline was implemented in R to allow for visualisation of reads alignments on specific transcripts. Unlike some other visualisation approaches, this algorithm uses genome-based alignments and bases visualizations on exons rather than transcripts. This obviates the need to use an isoform quantification tool and allows for manual interpretation of data to compare the potential for use of different regulatory mechanisms, such as alternative initiation and termination sites or alternative splicing (or readthrough or frameshifting into an extension ORF). I also identified an error in the Xtail ribosome profiling differential expression package and wrote a corrected version of the package source code that was incorporated into the public Xtail repository (Xiao et al., 2016).

The problem of ambiguously mapping reads was analysed in the specific context of ribosome profiling datasets. To my knowledge, this is the first investigation of the problem specifically in the context of the CDSome using RPF-length reads. Mammalian immune system genes were shown to be particularly affected by the presence of repeated sequences, as well as several cancer antigen, histone, and olfactory receptor genes. These analyses were run on the five genomes most commonly studied in previous profiling reports (human, mouse, yeast, *E. coli*, *Arabidopsis*), and each genome was shown to be potentially affected by the multi-mapping read problem. Lastly, I show how an existing toolkit, RSEM, can readily be adapted to RiboSeq workflows to allow for the retention of ambiguously mapping RPFs in a reasonably rigorous

manner (Li et al., 2009a). The custom scripts, pipelines, binaries and documentation developed during this work will be made publicly available via Github in the near future.

7.4 Challenges in mining RiboSeq datasets

While ribosome profiling has grown into more widespread usage since being introduced in 2009, much work remains to be done in the handling of the resulting data. The short nature of RPFs increases the probability of ambiguous mappings, which have yet to be dealt with in a rigorous and agreed upon manner. Very few profiling studies release all source code and scripts used in analyses to allow for exact replication of results. Even if these materials are made available, omission of critical details such as the version numbers of particular packages can lead to differences in results. While all published RiboSeq sequencing data are available via repositories such as GEO, there is currently no requirement or scientific consensus for releasing the associated software.

A handful of ribosome profiling specific packages for differential expression have been developed but have yet to be tested in comprehensive simulations. No package currently operates on the assumption of a Cauchy distribution, which, as the quotient of two negative binomial distributions, may represent a good model for genome-wide translation efficiencies. RNA binding proteins and ncRNA may produce RPF-like fragments that bias downstream analyses. All differential expression algorithms benefit from having more replicates available to estimate read count dispersions, though cost and time are limiting factors in producing large numbers of replicates.

Ligation, fragmentation, PCR and sequencing biases can all introduce distortion into profiling data that leads to unusual peaks or aberrant expression quantification. No published study (to my knowledge) has examined the role of transcript sequestration in stress granules or P-bodies and whether these species contribute to RPF and/or RNASeq read counts. Several studies have also noted that the use of translation inhibitors such as CHX can lead to re-distributions or distortions of RPF counts along transcripts, biasing analyses of short sequences or particular pause sites.

Discarding multi-mapping reads prohibits the analysis of certain genes (such as many mammalian immune system genes) while retention of these reads without proper controls can significantly inflate read count measurements.

7.5 Future Directions

Future work will focus on development of an automated pipeline to download, process, align and compare results with multi-mapping reads, to show what information was missed by discarding such reads, and in some cases, what results were biased by retention of multi-mapping reads, such as in the case of the public RPFdb database (Xie et al., 2016). The RSEM source code will be modified to refine the expectation maximization algorithm in order to first process RNASeq data and to then filter alignment indices such that the RiboSeq mapping space is constrained to regions for which there is transcriptomic support in a particular replicate. The HIV and DENV viral and cellular analyses will be repeated once biological (mycoplasma-free) replicates become available for these projects. The repeats will also be used to perform differential expression studies for cellular genes. Once initiation profiling (harringtonine treatment) is performed on the *Arabidopsis* datasets, a more detailed uORF utilization analysis will be performed. Use of Docker framework for compartmentalization of code will drastically improve reproducibility in RiboSeq bioinformatics analyses.

Ribosome profiling has revealed the widespread use of uORFs in many eukaryotic transcripts. There is much debate currently on the precise role of these elements, such as whether they are pre-dominantly regulatory in nature or encode micro-peptides that specify biological functions such as immunogenic identification or modulation of enzymatic reactions.

In our data and a previous study, the murine Slc35a4 5′ leader was shown to contain a 103 codon sequence that was highly conserved, had excellent RPF support and was potentially supported by mass spectrometry data. This UDP-sugar transporter protein may represent a rare example of a functionally bi-cistronic mammalian transcript. Further ribosome profiling (in various tissue types), targeted proteomics and genetic knockout studies will be useful in evaluating the role of such uORF sequences.

8 BIBLIOGRAPHY

- Aeschimann, F., Xiong, J., Arnold, A., Dieterich, C., and Großhans, H. (2015). Transcriptome-wide measurement of ribosomal occupancy by ribosome profiling. Methods *85*, 75–89.
- Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., et al. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. *44*, W3–W10.
- Ahlquist, P. (2006). Parallels among positive-strand RNA viruses, reverse-transcribing viruses and double-stranded RNA viruses. Nat. Rev. Microbiol. *4*, 371–382.
- Alatorre-Cobos, F., Cruz-Ramírez, A., Hayden, C.A., Pérez-Torres, C.A., Chauvin, A.L., Ibarra-Laclette, E., Alva-Cortés, E., Jorgensen, R.A., and Herrera-Estrella, L. (2012). Translational regulation of Arabidopsis XIPOTL1 is modulated by phosphocholine levels via the phylogenetically conserved upstream open reading frame 30. J. Exp. Bot. *63*, 5203–5221.
- Alcaraz-Estrada, S.L., Yocupicio-Monroy, M., and del Angel, R.M. (2010). Insights into dengue virus genome replication. Future Virol. *5*, 575–592.
- Aleshin, A.E., Shiryaev, S. a, Strongin, A.Y., and Liddington, R.C. (2007). Structural evidence for regulation and specificity of flaviviral proteases and evolution of the Flaviviridae fold. Protein Sci. *16*, 795–806.
- Allam, H., and Ali, N. (2010). Initiation factor eIF2-independent mode of c-Src mRNA translation occurs via an internal ribosome entry site. J. Biol. Chem. *285*, 5713–5725.
- Altelaar, A.F.M., Munoz, J., and Heck, A.J.R. (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet *14*, 35–48.
- Alvarez, D.E., Lodeiro, M.F., Ludueña, S.J., Pietrasanta, L.I., and Gamarnik, A. V (2005). Long-range RNA-RNA interactions circularize the dengue virus genome. J. Virol. *79*, 6631–6643.
- Alvarez, D.E., Filomatori, C. V., and Gamarnik, A. V. (2008). Functional analysis of dengue virus cyclization sequences located at the 5′ and 3′ UTRs. Virology *375*, 223–235.
- de Alwis, R., Williams, K.L., Schmid, M.A., Lai, C.-Y.Y., Patel, B., Smith, S.A., Crowe, J.E., Wang, W.-K.K., Harris, E., and de Silva, A.M. (2014). Dengue Viruses Are Enhanced by Distinct Populations of Serotype Cross-Reactive Antibodies in Human Immune Sera. PLoS Pathog. *10*, e1004386.
- Ambrosone, A., Costa, A., Leone, A., and Grillo, S. (2012). Beyond transcription: RNA-binding proteins as emerging regulators of plant response to environmental constraints. Plant Sci. *182*, 12–18.
- Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol *11*, R106.
- Anders, S., and Huber, W. (2012). Differential expression of RNA-Seq data at the gene level–the DESeq package. EMBL, Heidelberg, Ger.
- Anders, S., McCarthy, D.J., Chen, Y., Okoniewski, M., Smyth, G.K., Huber, W., and Robinson, M.D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat. Protoc. *8*, 1765–1786.
- Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

- Anderson, P., and Kedersha, N. (2006). RNA granules. J. Cell Biol. *172*, 803–808.
- Andreev, D.E., O'Connor, P.B., Fahey, C., Kenny, E.M., Terenin, I.M., Dmitriev, S.E., Cormican, P., Morris, D.W., Shatsky, I.N., and Baranov, P. V. (2015a). Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. Elife *4*, 1–21.
- Andreev, D.E., O'Connor, P.B.F., Zhdanov, A. V, Dmitriev, R.I., Shatsky, I.N., Papkovsky, D.B., and Baranov, P. V (2015b). Oxygen and glucose deprivation induces widespread alterations in mRNA translation within 20 minutes. Genome Biol *16*, 90.
- Andrews, S.J., and Rothnagel, J.A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. Nat. Rev. Genet. *15*, 193–204.
- Archer, S.K., Shirokikh, N.E., Beilharz, T.H., and Preiss, T. (2016). Dynamics of ribosome scanning and recycling revealed by translation complex profiling. Nature *535*, 570–574.
- Arciga-Reyes, L., Wootton, L., Kieffer, M., and Davies, B. (2006). UPF1 is required for nonsense-mediated mRNA decay (NMD) and RNAi in Arabidopsis. Plant J. *47*, 480–489.
- Arenz, S., Meydan, S., Starosta, A.L., Berninghausen, O., Beckmann, R., Vazquez-Laslop, N., and Wilson, D.N. (2014). Drug sensing by the ribosome induces translational arrest via active site perturbation. Mol. Cell *56*, 446–452.
- Arhel, N. (2010). Revisiting HIV-1 uncoating. Retrovirology *7*, 96.
- Artieri, C.G., and Fraser, H.B. (2014). Artieri CG, Fraser HB. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. Genome Res. 2014;24:2011-2021. Genome Res. *24*, 2011–2021.
- Artpradit, C., Robinson, L.N., Gavrilov, B.K., Rurak, T.T., Ruchirawat, M., and Sasisekharan, R. (2013). Recognition of heparan sulfate by clinical strains of dengue virus serotype 1 using recombinant subviral particles. Virus Res. *176*, 69–77.
- Asano, K. (2014). Why is start codon selection so precise in eukaryotes? Translation *2*, e28387.
- Atkins, J., Gesteland, R., Dinman, J.D., Connor, M.O., and Farabaugh, P.J. (2010). Recoding: expansion of decoding rules enriches gene expression. Nucleic Acids Mol. Biol. *24*, 221–247.
- Au, K.F., Underwood, J.G., Lee, L., and Wong, W.H. (2012). Improving PacBio Long Read Accuracy by Short Read Alignment. PLoS One *7*, e46679.
- Bäck, A.T., and Lundkvist, A. (2013). Dengue viruses - an overview. Infect. Ecol. Epidemiol. *3*, 1–21.
- Baden, L.R., Petersen, L.R., Jamieson, D.J., Powers, A.M., and Honein, M.A. (2016). Zika Virus. N. Engl. J. Med. *374*, 1552–1563.
- Baker, S.C., Yokomori, K., Dong, S., Carlisle, R., Gorbalenya, A.E., Koonin, E. V, and Lai, M.M. (1993). Identification of the catalytic sites of a papain-like cysteine proteinase of murine coronavirus. J. Virol. *67*, 6056–6063.
- Banerjee, S., Narayanan, K., Mizutani, T., and Makino, S. (2002). Murine coronavirus replication-induced p38 mitogen-activated protein kinase activation promotes interleukin-6 production and virus replication in cultured cells. J. Virol. *76*, 5937–5948.
- Barbosa, C., Peixeiro, I., and Romão, L. (2013). Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. PLoS One *9*, e1003529.
- Battiste, J.L., Mao, H., Rao, N.S., Tan, R., Muhandiram, D.R., Kay, L.E., Frankel, A.D., and Williamson, J.R. (1996). Alpha helix-RNA major groove recognition in an HIV-1 rev peptide-RRE RNA complex. Science. *273*, 1547–1551.

- Bazzini, A.A., Lee, M.T., and Giraldez, A.J. (2012). Ribosome Profiling Shows That miR-430 Reduces Translation Before Causing mRNA Decay in Zebrafish. Science. *336*, 233–237.
- Bechill, J., Chen, Z., Brewer, J.W., and Baker, S.C. (2008). Coronavirus infection modulates the unfolded protein response and mediates sustained translational repression. J. Virol. *82*, 4492–4501.
- Beier, H., and Grimm, M. (2001). Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. Nucleic Acids Res. *29*, 4767–4782.
- Belew, A.T., Meskauskas, A., and Musalgaonkar, S. (2014). Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway. Nature *512*, 265–269.
- Belle, A., Tanay, A., Bitincka, L., Shamir, R., and O'Shea, E.K. (2006). Quantification of protein half-lives in the budding yeast proteome. Proc. Natl. Acad. Sci. U. S. A. *103*, 13004–13009.
- Ben-Shem, A., Jenner, L., Yusupova, G., and Yusupov, M. (2010). Crystal structure of the eukaryotic ribosome. Science *330*, 1203–1209.
- Ben-Shem, A., Garreau de Loubresse, N., Melnikov, S., Jenner, L., Yusupova, G., and Yusupov, M. (2011). SOM: The structure of the eukaryotic ribosome at 3.0 Å resolution. Science *334*, 1524–1529.
- Benarroch, D., Selisko, B., Locatelli, G., and Maga, G. (2004). The RNA helicase, nucleotide 5'-triphosphatase, and RNA 5'-triphosphatase activities of Dengue virus protein NS3 are Mg2+-dependent and require a functional Walker B motif in the helicase catalytic core. Virology *328*, 208-218.
- Beniac, D.R., Andonov, A., Grudeski, E., and Booth, T.F. (2006). Architecture of the SARS coronavirus prefusion spike. Nat. Struct. Mol. Biol. *13*, 751–752.
- Benjamini, Y., and Hochberg, Y. (1995). Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B *57*, 289–300.
- Berg, J.M., Tymoczko, J.L., and Stryer, L. (2012). Biochemistry: International edition. In Biochemistry: International Edition, p. 1120.
- Beringer, M., and Rodnina, M. V. (2007). The Ribosomal Peptidyl Transferase. Mol. Cell *26*, 311–321.
- Bidet, K., and Garcia-Blanco, M.A. (2014). Flaviviral RNAs: weapons and targets in the war between virus and host. Biochem. J. *462*, 215–230.
- Bidet, K., Dadlani, D., and Garcia-Blanco, M.A. (2014). G3BP1, G3BP2 and CAPRIN1 Are Required for Translation of Interferon Stimulated mRNAs and Are Targeted by a Dengue Virus Non-coding RNA. PLoS Pathog. *10*, e1004242.
- Bieling, P., Beringer, M., Adio, S., and Rodnina, M. V (2006). Peptide bond formation does not involve acid-base catalysis by ribosomal residues. Nat. Struct. Mol. Biol. *13*, 423–428.
- Bischoff, L., Berninghausen, O., and Beckmann, R. (2014). Molecular basis for the ribosome functioning as an L-tryptophan sensor. Cell Rep. *9*, 469–475.
- Blaney, J.E., Sathe, N.S., Goddard, L., Hanson, C.T., Romero, T.A., Hanley, K.A., Murphy, B.R., and Whitehead, S.S. (2008). Dengue virus type 3 vaccine candidates generated by introduction of deletions in the 3' untranslated region (3'-UTR) or by exchange of the DENV-3 3'-UTR with that of DENV-4. Vaccine *26*, 817–828.
- Bollati, M., Alvarez, K., Assenberg, R., Baronti, C., Canard, B., Cook, S., Coutard, B., Decroly, E., de Lamballerie, X., Gould, E.A., et al. (2010). Structure and functionality in flavivirus NS-proteins: Perspectives for drug design. Antiviral Res. *87*, 125–148.

- Bourgon, R., Gentleman, R., and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. Proc. Natl. Acad. Sci. U. S. A. *107*, 9546–9551.
- Branco-Price, C., Kawaguchi, R., Ferreira, R.B., Bailey-Serres, J., and Bailey-Serres, J. (2005). Genome-wide Analysis of Transcript Abundance and Translation in Arabidopsis Seedlings Subjected to Oxygen Deprivation. Ann. Bot. *96*, 647–660.
- Brar, G. a, and Weissman, J.S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. Nat. Rev. Mol. Cell Biol. *16*, 651–664.
- Brar, G.A., Yassour, M., Friedman, N., Regev, A., Ingolia, N.T., and Weissman, J.S. (2012). High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling. Science. *335*, 552–557.
- Bravo, J., Aguilar-Henonin, L., Olmedo, G., and Guzmán, P. (2005). Four distinct classes of proteins as interaction partners of the PABC domain of Arabidopsis thaliana Poly(A)-binding proteins. Mol. Genet. Genomics *272*, 651–665.
- Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat Biotech *34*, 525–527.
- Brierley, I., and Dos Ramos, F.J. (2006). Programmed ribosomal frameshifting in HIV-1 and the SARS-CoV. Virus Res *119*, 29–42.
- Brierley, I., Boursnell, M.E., Binns, M.M., Bilimoria, B., Blok, V.C., Brown, T.D., and Inglis, S.C. (1987). An efficient ribosomal frame-shifting signal in the polymerase-encoding region of the coronavirus IBV. EMBO J. *6*, 3779–3785.
- Brown, J.D., and Ryan, M.D. (2010). Ribosome ``Skipping'': ``Stop-Carry On'' or ``StopGo'' Translation. In Recoding: Expansion of Decoding Rules Enriches Gene Expression, J.F. Atkins, and R.F. Gesteland, eds. (New York, NY: Springer New York), pp. 101–121.
- Browning, K.S., and Bailey-Serres, J. (2015). Mechanism of Cytoplasmic mRNA Translation. Arab. B. *13*, e0176.
- Brubaker, S.W., Gauthier, A.E., Mills, E.W., Ingolia, N.T., and Kagan, J.C. (2014). A bicistronic MAVS transcript highlights a class of truncated variants in antiviral immunity. Cell *156*, 800–811.
- Buermans, H.P.J., and den Dunnen, J.T. (2014). Next generation sequencing technology: Advances and applications. Biochim. Biophys. Acta - Mol. Basis Dis. *1842*, 1932–1941.
- Buffington, S. a, Huang, W., and Costa-Mattioli, M. (2014). Translational Control in Synaptic Plasticity and Cognitive Dysfunction. Annu. Rev. Neurosci. *37*, 17–38.
- Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics *11*, 94.
- Buzon, V., Natrajan, G., Schibli, D., Campelo, F., Kozlov, M.M., and Weissenhorn, W. (2010). Crystal structure of HIV-1 gp41 including both fusion peptide and membrane proximal external regions. PLoS Pathog. *6*, 1–7.
- Cabrera-Hernandez, A., Thepparit, C., Suksanpaisan, L., and Smith, D.R. (2007). Dengue virus entry into liver (HepG2) cells is independent of hsp90 and hsp70. J. Med. Virol. *79*, 386–392.
- Calfon, M., Zeng, H., Urano, F., Till, J.H., Hubbard, S.R., Harding, H.P., Clark, S.G., and Ron, D. (2002). IRE1 couples endoplasmic reticulum load to secretory capacity by processing the XBP-1 mRNA. Nature *415*, 92–96.

- Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., and Ohler, U. (2015). Detecting actively translated open reading frames in ribosome profiling data. Nat. Methods *13*, 1–9.
- Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc. Natl. Acad. Sci. U. S. A. *106*, 7507–7512.
- Cambridge, S.B., Gnad, F., Nguyen, C., Bermejo, J.L., Krüger, M., and Mann, M. (2011). Systems-wide proteomic analysis in mammalian cells reveals conserved, functional protein turnover. J. Proteome Res. *10*, 5275–5284.
- Carberry, S.E., Darzynkiewicz, E., and Goss, D.J. (1991). A comparison of the binding of methylated cap analogues to wheat germ protein synthesis initiation factors 4F and (iso)4F. Biochemistry *30*, 1624–1627.
- Carroll, A.J. (2013). The Arabidopsis Cytosolic Ribosomal Proteome: From form to Function. Front. Plant Sci. *4*, 32.
- Carroll, A.J., Heazlewood, J.L., Ito, J., and Millar, A.H. (2008). Analysis of the Arabidopsis cytosolic ribosome proteome provides detailed insights into its components and their post-translational modification. Mol. Cell. Proteomics *7*, 347–369.
- Carvalho, S.E.S., Martin, D.P., Oliveira, L.M., Ribeiro, B.M., and Nagata, T. (2010). Comparative analysis of American Dengue virus type 1 full-genome sequences. Virus Genes *40*, 60–66.
- Carvunis, A.-R. (2012). Proto-genes and de novo gene birth. Nature *487*, 370–374.
- Cassan, E., Arigon-Chifolleau, A.-M., Mesnard, J.-M., Gross, A., and Gascuel, O. (2016). Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic. Proc. Natl. Acad. Sci. .
- Cavanagh, D. (2007). Coronavirus avian infectious bronchitis virus. Vet. Res. *38*, 281–297.
- Cen, S., Niu, M., Saadatmand, J., Guo, F., Huang, Y., Nabel, G.J., and Kleiman, L. (2004). Incorporation of pol into human immunodeficiency virus type 1 Gag virus-like particles occurs independently of the upstream Gag domain in Gag-pol. J Virol *78*, 1042–1049.
- Chakraborty, R., Baek, J.H., Bae, E.Y., Kim, W.-Y., Lee, S.Y., and Kim, M.G. (2016). Comparison and contrast of plant, yeast, and mammalian ER stress and UPR. Appl. Biol. Chem. *59*, 337–347.
- Chamanian, M., Purzycka, K.J., Wille, P.T., Ha, J.S., McDonald, D., Gao, Y., Le Grice, S.F.J., and Arts, E.J. (2013). A cis-acting element in retroviral genomic RNA links Gag-Pol ribosomal frameshifting to selective viral RNA encapsidation. Cell Host Microbe *13*, 181–192.
- Chan, J.F.W., Lau, S.K.P., To, K.K.W., Cheng, V.C.C., Woo, P.C.Y., and Yue, K.Y. (2015). Middle East Respiratory syndrome coronavirus: Another zoonotic betacoronavirus causing SARS-like disease. Clin. Microbiol. Rev. *28*, 465–522.
- Chang, I.E. Al (2005). Proteomic Characterization of Evolutionarily Conserved and Variable Proteins of Arabidopsis. Society *137*, 848–862.
- Chapman, E.G., Costantino, D.A., Rabe, J.L., Moon, S.L., Wilusz, J., Nix, J.C., and Kieft, J.S. (2014a). The Structural Basis of Pathogenic Subgenomic Flavivirus RNA (sfRNA) Production. Science. *344*, 307–310.
- Chapman, E.G., Moon, S.L., Wilusz, J., and Kieft, J.S. (2014b). RNA structures that resist degradation by Xrn1 produce a pathogenic dengue virus RNA. Elife *2014*, e01892.
- Charley, P.A., and Wilusz, J. (2014). Sponging of cellular proteins by viral RNAs. Curr. Opin. Virol. *9*, 14–18.

- Charley, P.A., and Wilusz, J. (2016). Standing your ground to exoribonucleases: Function of Flavivirus long non-coding RNAs. Virus Res. *212*, 70–77.
- Charneski, C.A., and Hurst, L.D. (2013). Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. PLoS Biol. *11*, e1001508.
- Chekulaeva, M., and Landthaler, M. (2016). Eyes on Translation. Mol. Cell *63*, 918–925.
- Chen, L.K., Liao, C.L., Lin, C.G., Lai, S.C., Liu, C.I., Ma, S.H., Huang, Y.Y., and Lin, Y.L. (1996). Persistence of Japanese encephalitis virus is associated with abnormal expression of the nonstructural protein NS1 in host cells. Virology *217*, 220–229.
- Chen, X.-G., Jiang, X., Gu, J., Xu, M., Wu, Y., Deng, Y., Zhang, C., Bonizzoni, M., Dermauw, W., Vontas, J., et al. (2015a). Genome sequence of the Asian Tiger mosquito, Aedes albopictus, reveals insights into its biology, genetics, and evolution. Proc. Natl. Acad. Sci. U. S. A. *112*, E5907-5915.
- Chen, Y., Savinov, S.N., Mielech, A.M., Cao, T., Baker, S.C., and Mesecar, A.D. (2015b). X-ray Structural and Functional Studies of the Three Tandemly-Linked Domains of Nsp3 from Murine Hepatitis Virus Reveal Conserved Functions. J. Biol. Chem. *3*, jbc.M115.662130.
- Chen, Y.G., Yuan, K., Zhang, Z.Z., Yuan, F.H., Weng, S.P., Yue, H.T., He, J.G., and Chen, Y.H. (2016). Identification and functional characterization of a solute carrier family 15, member 4 gene in Litopenaeus vannamei. Dev. Comp. Immunol. *57*, 57–66.
- Chen, Y.J., Tan, B.C.M., Cheng, Y.Y., Chen, J.S., and Lee, S.C. (2009). Differential regulation of CHOP translation by phosphorylated eIF4E under stress conditions. Nucleic Acids Res. *38*, 764–777.
- Chinnusamy, V., Zhu, J., and Zhu, J.-K. (2007). Cold stress regulation of gene expression in plants. Trends Plant Sci. *12*, 444–451.
- Chiu, W.-W., Kinney, R.M., and Dreher, T.W. (2005). Control of translation by the 5'- and 3'-terminal regions of the dengue virus genome. J. Virol. *79*, 8303–8315.
- Chotewutmontri, P., and Barkan, A. (2016). Dynamics of Chloroplast Translation during Chloroplast Differentiation in Maize. PLOS Genet. *12*, e1006106.
- Chotiwan, N., Roehrig, J.T., Schlesinger, J.J., Blair, C.D., and Huang, C.Y.H. (2014). Molecular determinants of dengue virus 2 envelope protein important for virus entry in FcγRIIA-mediated antibody-dependent enhancement of infection. Virology *456–457*, 238–246.
- Chu, J., Cargnello, M., Topisirovic, I., and Pelletier, J. (2016). Translation Initiation Factors: Reprogramming Protein Synthesis in Cancer. Trends Cell Biol. *12*, 1–16.
- Chua, J.J.E., Ng, M.M.L., and Chow, V.T.K. (2004). The non-structural 3 (NS3) protein of dengue virus type 2 interacts with human nuclear receptor binding protein and is associated with alterations in membrane structure. Virus Res. *102*, 151–163.
- Chung, B.Y., Hardcastle, T.J., Jones, J.D., Irigoyen, N., Firth, A.E., Baulcombe, D.C., and Brierley, I. (2015). The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis. RNA *21*, 1731–1745.
- Chung, D., Kuan, P.F., Li, B., Sanalkumar, R., Liang, K., Bresnick, E.H., Dewey, C., and Keleş, S. (2011). Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-seq data. PLoS Comput. Biol. *7*, e1002111.
- Clapham, P.R., and McKnight, Á. (2001). HIV-1 receptors and cell tropism. Br. Med. Bull. *58*, 43–59.
- Clarke, B.D., Roby, J.A., Slonchak, A., and Khromykh, A.A. (2015). Functional non-coding RNAs derived from the flavivirus 3' untranslated region. Virus Res. *206*, 53–61.

- Clyde, K., and Harris, E. (2006). RNA secondary structure in the coding region of dengue virus type 2 directs translation start codon selection and is required for viral replication. J. Virol. *80*, 2170–2182.
- Coakley, E., Petropoulos, C.J., and Whitcomb, J.M. (2005). Assessing chemokine co-receptor usage in HIV. Curr. Opin. Infect. Dis. *18*, 9–15.
- Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.
- Coleman, C.M., and Frieman, M.B. (2014). Coronaviruses: important emerging human pathogens. J. Virol. *88*, 5209–5212.
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., et al. (2016). A survey of best practices for RNA-seq data analysis. Genome Biol. *17*, 13.
- Cridge, A.G., Major, L.L., Mahagaonkar, A.A., Poole, E.S., Isaksson, L.A., and Tate, W.P. (2006). Comparison of characteristics and function of translation termination signals between and within prokaryotic and eukaryotic organisms. Nucleic Acids Res. *34*, 1959–1973.
- Csárdi, G., Franks, A., Choi, D.S., Airoldi, E.M., and Drummond, D.A. (2015). Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast. PLOS Genet. *11*, e1005206.
- Csibra, E., Brierley, I., and Irigoyen, N. (2014). Modulation of stop codon read-through efficiency and its effect on the replication of murine leukemia virus. J Virol *88*, 10364–10376.
- Curran, J., and Kolakofsky, D. (1988). Ribosomal initiation from an ACG codon in the Sendai virus P/C mRNA. EMBO J. *7*, 245–251.
- D'Onofrio, G., Jabbari, K., Musto, H., and Bernardi, G. (1999). The correlation of protein hydropathy with the base composition of coding sequences. Gene *238*, 3–14.
- Dallas, D.C., Guerrero, A., Parker, E.A., Robinson, R.C., Gan, J., German, J.B., Barile, D., and Lebrilla, C.B. (2015). Current peptidomics: Applications, purification, identification, quantification, and functional analysis. Proteomics *15*, 1026–1038.
- Dalrymple, N.A., Cimica, V., and Mackow, E.R. (2015). Dengue virus NS proteins inhibit RIG-I/MAVS signaling by blocking TBK1/IRF3 phosphorylation: Dengue virus serotype 1 NS4A is a unique interferon-regulating virulence determinant. MBio *6*, 1–12.
- Dana, A., and Tuller, T. (2012). Determinants of Translation Elongation Speed and Ribosomal Profiling Biases in Mouse Embryonic Stem Cells. PLoS Comput. Biol. *8*.
- Das, S., Laxminarayana, S.V., Chandra, N., Ravi, V., and Desai, A. (2009). Heat shock protein 70 on Neuro2a cells is a putative receptor for Japanese encephalitis virus. Virology *385*, 47–57.
- Daugherty, M.D., Booth, D.S., Jayaraman, B., Cheng, Y., and Frankel, A.D. (2010). HIV Rev response element (RRE) directs assembly of the Rev homooligomer into discrete asymmetric complexes. Proc. Natl. Acad. Sci. U. S. A. *107*, 12481–12486.
- Davidson, A.D. (2009). New Insights into Flavivirus Nonstructural Protein 5.
- Davis, A.R., Gohara, D.W., and Yap, M.-N.F. (2014). Sequence selectivity of macrolide-induced translational attenuation. Proc. Natl Acad. Sci. USA *111*, 15379–15384.

- Dejima, K., Murata, D., Mizuguchi, S., Nomura, K.H., Gengyo-Ando, K., Mitani, S., Kamiyama, S., Nishihara, S., and Nomura, K. (2009). The ortholog of human solute carrier family 35 member B1 (UDP-galactose transporter-related protein 1) is involved in maintenance of ER homeostasis and essential for larval development in Caenorhabditis elegans. FASEB J. *23*, 2215–2225.
- Delhomme, N., Padioleau, I., Furlong, E.E., and Steinmetz, L.M. (2012). easyRNASeq: A bioconductor package for processing RNA-Seq data. Bioinformatics *28*, 2532–2533.
- Dever, T.E., and Green, R. (2012). The elongation, termination, and recycling phases of translation in eukaryotes. Cold Spring Harb. Perspect. Biol. *4*, 1–16.
- Diament, A., and Tuller, T. (2016). Estimation of ribosome profiling performance and reproducibility at various levels of resolution. Biol. Direct *11*, 1–12.
- Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C., and Assmann, S.M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature *505*, 696–700.
- Dinger, M.E., Pang, K.C., Mercer, T.R., and Mattick, J.S. (2008). Differentiating protein-coding and noncoding RNA: Challenges and ambiguities. PLoS Comput. Biol. *4*.
- Dinkova, T.D., Zepeda, H., Martínez-Salas, E., Martínez, L.M., Nieto-Sotelo, J., and Sánchez De Jiménez, E. (2005). Cap-independent translation of maize Hsp101. Plant J. *41*, 722–731.
- Dinman, J.D. (2009). The eukaryotic ribosome: Current status and challenges. J. Biol. Chem. *284*, 11761–11765.
- Dinman, J.D., and Berry, M.J. (2007). 22 Regulation of Termination and Recoding. Cold Spring Harb. Monogr. Arch. *48*, 625–654.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.
- Doerfel, L.K., Wohlgemuth, I., Kothe, C., Peske, F., Urlaub, H., and Rodnina, M. V. (2013). EF-P Is Essential for Rapid Synthesis of Proteins Containing Consecutive Proline Residues. Science. *339*, 85–88.
- Dong, H., Ray, D., Ren, S., Zhang, B., Puig-Basagoiti, F., Takagi, Y., Ho, C.K., Li, H., and Shi, P.Y. (2007). Distinct RNA elements confer specificity to flavivirus RNA cap methylation events. J Virol *81*, 4412–4421.
- Donnelly, M.L.L., Luke, G., Mehrotra, A., Li, X., Hughes, L.E., Gani, D., and Ryan, M.D. (2001). Analysis of the aphthovirus 2A/2B polyprotein “cleavage” mechanism indicates not a proteolytic reaction, but a novel translational effect: A putative ribosomal “skip.” J. Gen. Virol. *82*, 1013–1025.
- Doyon, L., Payant, C., Brakier-Gingras, L., and Lamarre, D. (1998). Novel Gag-Pol frameshift site in human immunodeficiency virus type 1 variants resistant to protease inhibitors. J. Virol. *72*, 6146–6150.
- Dulude, D., Berchiche, Y.A., Gendron, K., Brakier-Gingras, L., and Heveker, N. (2006). Decreasing the frameshift efficiency translates into an equivalent reduction of the replication of the human immunodeficiency virus type 1. Virology *345*, 127–136.
- Dunn, J.G., Foo, C.K., Belletier, N.G., Gavis, E.R., and Weissman, J.S. (2013). Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. Elife *2013*.

- Dutt, S., Parkash, J., Mehra, R., Sharma, N., Singh, B., Raigond, P., Joshi, A., Chopra, S., and Singh, B.P. (2015). Translation initiation in plants: roles and implications beyond protein synthesis. Biol. Plant. *59*, 401–412.
- Echevarría-Zomeño, S., Yángüez, E., Fernández-Bautista, N., Castro-Sanz, A.B., Ferrando, A., and Castellano, M.M. (2013). Regulation of translation initiation under biotic and abiotic stresses. Int. J. Mol. Sci. *14*, 4670–4683.
- Edgil, D., Diamond, M.S., Holden, K.L., Paranjape, S.M., and Harris, E. (2003). Translation efficiency determines differences in cellular infection among dengue virus type 2 strains. Virology *317*, 275–290.
- Emara, M.M., and Brinton, M.A. (2007). Interaction of TIA-1/TIAR with West Nile and dengue virus products in infected cells interferes with stress granule formation and processing body assembly. Proc. Natl. Acad. Sci. U. S. A. *104*, 9041–9046.
- Epple, L.M., Dodd, R.D., Merz, A.L., Dechkovskaia, A.M., Herring, M., Winston, B.A., Lencioni, A.M., Russell, R.L., Madsen, H., Nega, M., et al. (2013). Induction of the Unfolded Protein Response Drives Enhanced Metabolism and Chemoresistance in Glioma Cells. PLoS One *8*, e73267.
- Erbel, P., Schiering, N., D'Arcy, A., Renatus, M., Kroemer, M., Lim, S.P., Yin, Z., Keller, T.H., Vasudevan, S.G., and Hommel, U. (2006). Structural basis for the activation of flaviviral NS3 proteases from dengue and West Nile virus. Nat. Struct. Mol. Biol. *13*, 372–373.
- Ernst, H., and Shatkin, a J. (1985). Reovirus hemagglutinin mRNA codes for two polypeptides in overlapping reading frames. Proc. Natl. Acad. Sci. U. S. A. *82*, 48–52.
- Falgout, B., and Markoff, L. (1995). Evidence that flavivirus NS1-NS2A cleavage is mediated by a membrane-bound host protease in the endoplasmic reticulum. J. Virol. *69*, 7232–7243.
- Farabaugh, P.J. (2000). Translational frameshifting: implications for the mechanism of translational frame maintenance. Prog. Nucleic Acid Res. Mol. Biol. *64*, 131–170.
- De Felipe, P., Hughes, L.E., Ryan, M.D., and Brown, J.D. (2003). Co-translational, intraribosomal cleavage of polypeptides by the foot-and-mouth disease virus 2A peptide. J. Biol. Chem. *278*, 11441–11448.
- Feng, Y., Zhang, Y., Ying, C., Wang, D., and Du, C. (2015). Nanopore-based fourth-generation DNA sequencing technology. Genomics, Proteomics Bioinforma. *13*, 4–16.
- Fernandes, J., Jayaraman, B., and Frankel, A. (2012). The HIV-1 Rev response element: an RNA scaffold that directs the cooperative assembly of a homo-oligomeric ribonucleoprotein complex. RNA Biol *9*, 6–11.
- Ferreira, S., Hjernø, K., Larsen, M., Wingsle, G., Larsen, P., Fey, S., Roepstorff, P., and Salomé Pais, M. (2006). Proteome profiling of Populus euphratica Oliv. upon heat stress. Ann. Bot. *98*, 361–377.
- Fields, A.P., Rodriguez, E.H., Jovanovic, M., Stern-Ginossar, N., Haas, B.J., Mertins, P., Raychowdhury, R., Hacohen, N., Carr, S.A., Ingolia, N.T., et al. (2015). A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. Mol. Cell *60*, 816–827.
- Fields, B.N., Knipe, D.M., and Howley, P.M. (2013). Fields Virology, 6th Edition.
- Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W.K., and Mockler, T.C. (2010). Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res. *20*, 45–58.

- Filomatori, C. V., Lodeiro, M.F., Alvarez, D.E., Samsa, M.M., Pietrasanta, L., and Gamarnik, A. V. (2006). A 5′ RNA element promotes dengue virus RNA synthesis on a circular genome. Genes Dev. *20*, 2238–2249.
- Filomatori, C. V., Iglesias, N.G., Villordo, S.M., Alvarez, D.E., and Gamarnik, A. V. (2011). RNA sequences and structures required for the recruitment and activity of the dengue virus polymerase. J. Biol. Chem. *286*, 6929–6939.
- Firth, A.E., and Atkins, J.F. (2009). A conserved predicted pseudoknot in the NS2A-encoding sequence of West Nile and Japanese encephalitis flaviviruses suggests NS1′ may derive from ribosomal frameshifting. Virol. J. *6*, 14.
- Firth, A.E., and Brierley, I. (2012). Non-canonical translation in RNA viruses. J. Gen. Virol. *93*, 1385–1409.
- Fischer, F., Peng, D., Hingley, S.T., Weiss, S.R., and Masters, P.S. (1997). The internal open reading frame within the nucleocapsid gene of mouse hepatitis virus encodes a structural protein that is not essential for viral replication. J. Virol. *71*, 996–1003.
- Flower, a M., and McHenry, C.S. (1990). The gamma subunit of DNA polymerase III holoenzyme of Escherichia coli is produced by ribosomal frameshifting. Proc. Natl. Acad. Sci. U. S. A. *87*, 3713–3717.
- Foley, B., Leitner, T., Apetrei, C., Hahn, B., Mizrachi, I., Mullins, J., A, R., S, W., and B, K. (2013). HIV Sequence Compendium 2013.
- Fournier, C.T. (2012). Amino termini of many yeast proteins map to downstream start codons. J. Proteome Res. *11*, 5712–5719.
- Fowler, S., and Thomashow, M.F. (2002). Arabidopsis transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. Plant Cell *14*, 1675–1690.
- Freed, E.O., and Martin, M.A. (2013). Human Immunodeficiency Viruses: Replication. In Fields Virology. 6th Edition. II., H.P. Knipe DM, ed. (Wolters Kluwer, Lippincott Williams & Wilkens), pp. 1503–1560.
- Fritsch, C. (2012). Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Res. *22*, 2208–2218.
- Fuller, F., Bhown, A.S., and Bishop, D.H.L. (1983). Bunyavirus nucleoprotein, N, and a non-structural protein, NS(s), are coded by overlapping reading frames in the S RNA. J. Gen. Virol. *64*, 1705–1714.
- Fung, T.S., and Liu, D.X. (2014). Coronavirus infection, ER stress, apoptosis and innate immunity. Frontiers in Microbiology *5*, 296.
- Fung, T., Liao, Y., and Liu, D. (2016). Regulation of Stress Responses and Translational Control by Coronavirus. Viruses *8*, 184.
- Futterer, J., Kiss-Laszlo, Z., and Hohn, T. (1993). Nonlinear ribosome migration on cauliflower mosaic virus 35S RNA. Cell *73*, 789–802.
- Gadlage, M.J., and Denison, M.R. (2010). Exchange of the Coronavirus Replicase Polyprotein Cleavage Sites Alters Protease Specificity and Processing. J. Virol. *84*, 6894–6898.
- Galiano, V., Garcia-Valtanen, P., Micol, V., and Encinar, J.A. (2016). Looking for inhibitors of the dengue virus NS5 RNA-dependent RNA-polymerase using a molecular docking approach. Drug Des. Devel. Ther. *10*, 3163–3181.
- Gao, H., Brandizzi, F., Benning, C., and Larkin, R.M. (2008). A membrane-tethered transcription factor defines a branch of the heat stress response in Arabidopsis thaliana. Proc. Natl. Acad. Sci. U. S. A. *105*, 16398–16403.

- García-Montalvo, B.M., Medina, F., and Del Angel, R.M. (2004). La protein binds to NS5 and NS3 and to the 5′ and 3′ ends of Dengue 4 virus RNA. Virus Res. *102*, 141–150.
- Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., and Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. *31*, 3784–3788.
- Gebauer, F., Preiss, T., and Hentze, M.W. (2012). From cis-regulatory elements to complex RNPs and back. Cold Spring Harb. Perspect. Biol. *4*, 1–14.
- Gebhard, L.G., Filomatori, C. V., and Gamarnik, A. V. (2011). Functional RNA elements in the dengue virus genome. Viruses *3*, 1739–1756.
- Gerashchenko, M. V., and Gladyshev, V.N. (2014). Translation inhibitors cause abnormalities in ribosome profiling experiments. Nucleic Acids Res. *42*, e134.
- Gerashchenko, M. V, Lobanov, A. V, and Gladyshev, V.N. (2012). Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. Proc. Natl. Acad. Sci. U. S. A. *109*, 17394–17399.
- Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737–741.
- Giavalisco, P., Wilson, D., Kreitler, T., Lehrach, H., Klose, J., Gobom, J., and Fucini, P. (2005). High heterogeneity within the ribosomal proteins of the Arabidopsis thaliana 80S ribosome. Plant Mol. Biol. *57*, 577–591.
- Giedroc, D.P., and Cornish, P. V. (2009). Frameshifting RNA pseudoknots: Structure and mechanism. Virus Res. *139*, 193–208.
- Gilks, N., Kedersha, N., Ayodele, M., Shen, L., Stoecklin, G., Dember, L.M., and Anderson, P. (2004). Stress granule assembly is mediated by prion-like aggregation of TIA-1. Mol. Biol. Cell *15*, 5383–5398.
- Girnary, R., King, L., Robinson, L., Elston, R., and Brierley, I. (2007). Structure-function analysis of the ribosomal frameshifting signal of two human immunodeficiency virus type 1 isolates with increased resistance to viral protease inhibitors. J. Gen. Virol. *88*, 226–235.
- Goff, S.P. (2013). Retroviridae. In Fields Virology. 6th Edition. II., H.P. Knipe DM, ed. (Wolters Kluwer, Lippincott Williams & Wilkens), pp. 1424–1473.
- Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet *17*, 333–351.
- Goss, D.J., and Kleiman, F.E. (2013). Poly(A) binding proteins: Are they all created equal? Wiley Interdiscip. Rev. RNA *4*, 167–179.
- Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.
- Gu, Y., Harley, I.T.W., Henderson, L.B., Aronow, B.J., Vietor, I., Huber, L.A., Harley, J.B., Kilpatrick, J.R., Langefeld, C.D., Williams, A.H., et al. (2009). Identification of IFRD1 as a modifier gene for cystic fibrosis lung disease. Nature *458*, 1039–1042.
- Gualerzi, C.O., and Pon, C.L. (2015). Initiation of mRNA translation in bacteria: Structural and dynamic aspects. Cell. Mol. Life Sci. *72*, 4341–4367.
- Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835–840.

- Gutsche, I., Coulibaly, F., Voss, J.E., Salmon, J., d'Alayer, J., Ermonval, M., Larquet, E., Charneau, P., Krey, T., Mégret, F., et al. (2011). Secreted dengue virus nonstructural protein NS1 is an atypical barrel-shaped high-density lipoprotein. Proc. Natl. Acad. Sci. U. S. A. *108*, 8003–8008.
- Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S., and Lander, E.S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell *154*, 240–251.
- Guydosh, N.R., and Green, R. (2014). Dom34 rescues ribosomes in 3[prime] untranslated regions. Cell *156*, 950–962.
- de Haan, C.A.M., Stadler, K., Godeke, G.-J., Bosch, B.J., and Rottier, P.J.M. (2004). Cleavage inhibition of the murine coronavirus spike protein by a furin-like enzyme affects cell-cell but not virus-cell fusion. J. Virol. *78*, 6048–6054.
- von der Haar, T. (2008). A quantitative estimation of the global translational activity in logarithmically growing yeast cells. BMC Syst. Biol. *2*, 87.
- Häcker, H., Tseng, P.-H., and Karin, M. (2011). Expanding TRAF function: TRAF3 as a tri-faced immune regulator. Nat. Rev. Immunol. *11*, 457–468.
- Hanfrey, C., Franceschetti, M., Mayer, M.J., Illingworth, C., and Michael, A.J. (2002). Abrogation of upstream open reading frame-mediated translational control of a plant S-adenosylmethionine decarboxylase results in polyamine disruption and growth perturbations. J. Biol. Chem. *277*, 44131–44139.
- Hannah, M.A., Heyer, A.G., and Hincha, D.K. (2005). A global survey of gene regulation during cold acclimation in Arabidopsis thaliana. PLoS Genet. *1*, 0179–0196.
- Hansen, K.D., Brenner, S.E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. *38*, e131–e131.
- Hanson, J., Hanssen, M., Wiese, A., Hendriks, M.M.W.B., and Smeekens, S. (2008). The sucrose regulated transcription factor bZIP11 affects amino acid metabolism by regulating the expression of asparagine synthetase 1 and proline dehydrogenase 2. Plant J. *53*, 935–949.
- Hardcastle, T.J., and Kelly, K.A. (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics *11*, 422.
- Hatano, M., Umemura, M., Kimura, N., Yamazaki, T., Takeda, H., Nakano, H., Takahashi, S., and Takahashi, Y. (2013). The 5'-untranslated region regulates ATF5 mRNA stability via nonsense-mediated mRNA decay in response to environmental stress. FEBS J. *280*, 4693–4707.
- Hayden, C. a, and Jorgensen, R. a (2007). Identification of novel conserved peptide uORF homology groups in Arabidopsis and rice reveals ancient eukaryotic origin of select groups and preferential association with transcription factor-encoding genes. BMC Biol. *5*, 32.
- Haydon, M.J., Mielczarek, O., Robertson, F.C., Hubbard, K.E., and Webb, A.A.R. (2013). Photosynthetic entrainment of the Arabidopsis thaliana circadian clock. Nature *502*, 689–692.
- He, B. (2006). Viruses, endoplasmic reticulum stress, and interferon responses. Cell Death Differ. *13*, 393–403.
- Heaton, N.S., Moshkina, N., Fenouil, R., Gardner, T.J., Aguirre, S., Shah, P.S., Zhao, N., Manganaro, L., Hultquist, J.F., Noel, J., et al. (2016). Targeting Viral Proteostasis Limits Influenza Virus, HIV, and Dengue Virus Infection. Immunity *44*, 46–58.

- Heggie, L., and Halliday, K.J. (2005). The highs and lows of plant life: Temperature and light interactions in development. Int. J. Dev. Biol. *49*, 675–687.
- Heguy, A. (1997). Inhibition of the HIV Rev transactivator : a new target for therapeutic intervention. Front. Biosci. *2*, d283-97.
- Hemmila, E., Turbide, C., Olson, M., Jothy, S., Holmes, K. V, and Beauchemin, N. (2004). Ceacam1a-/- mice are completely resistant to infection by murine coronavirus mouse hepatitis virus A59. J. Virol. *78*, 10156–10165.
- Hernández, G., and Vazquez-Pianzola, P. (2005). Functional diversity of the eukaryotic translation initiation factors belonging to eIF4 families. Mech. Dev. *122*, 865–876.
- Hett, E.C., Kyne, R.E., Gopalsamy, A., Tones, M.A., Xu, H., Thio, G.L., Nolan, E., and Jones, L.H. (2016). Selectivity determination of a small molecule chemical probe using protein microarray and affinity capture techniques. ACS Comb. Sci. acscombsci.6b00089.
- Hill, M., Tachedjian, G., and Mak, J. (2005). The packaging and maturation of the HIV-1 Pol proteins. Curr. HIV Res. *3*, 73–85.
- Hinnebusch, A.G. (2005). Translational regulation of GCN4 and the general amino acid control of yeast. Annu. Rev. Microbiol. *59*, 407–450.
- Hinnebusch, A.G. (2014). The scanning mechanism of eukaryotic translation initiation. Annu. Rev. Biochem. *83*, 779–812.
- Hirai, A., Ohtsuka, N., Ikeda, T., Taniguchi, R., Blau, D., Nakagaki, K., Miura, H.S., Ami, Y., Yamada, Y.K., Itohara, S., et al. (2010). Role of mouse hepatitis virus (MHV) receptor murine CEACAM1 in the resistance of mice to MHV infection: studies of mice with chimeric mCEACAM1a and mCEACAM1b. J. Virol. *84*, 6654–6666.
- Hiramatsu, N., Joseph, V.T., and Lin, J.H. (2011). Monitoring and manipulating mammalian unfolded protein response. Methods Enzymol. *491*, 183–198.
- Hirsch, C.D., Springer, N.M., and Hirsch, C.N. (2015). Genomic limitations to RNA sequencing expression profiling. Plant J. *84*, 491–503.
- Hirsh, D. (1971). Tryptophan transfer RNA as the UGA suppressor. J. Mol. Biol. *58*, 439–458.
- Hofer, U. (2013). Viral evolution: Fooling the coronavirus proofreading machinery. Nat Rev Micro *11*, 662–663.
- Holden, K.L., and Harris, E. (2004). Enhancement of dengue virus translation: Role of the 3' untranslated region and the terminal 3' stem-loop domain. Virology *329*, 119–133.
- Holden, K.L., Stein, D.A., Pierson, T.C., Ahmed, A.A., Clyde, K., Iversen, P.L., and Harris, E. (2006). Inhibition of dengue virus translation and RNA synthesis by a morpholino oligomer targeted to the top of the terminal 3' stem-loop structure. Virology *344*, 439–452.
- Hollien, J., and Weissman, J.S. (2006). Decay of endoplasmic reticulum-localized mRNAs during the unfolded protein response. Science. *313*, 104–107.
- Hollien, J., Lin, J.H., Li, H., Stevens, N., Walter, P., and Weissman, J.S. (2009). Regulated Ire1-dependent decay of messenger RNAs in mammalian cells. J. Cell Biol. *186*, 323–331.
- Holmes, E.C. (2009). RNA virus genomics: A world of possibilities. J. Clin. Invest. *119*, 2488–2495.
- Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., Young, R.A., Lee, T.I., et al. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. Cell *95*, 717–728.

- Hori, N., Denesyuk, N.A., and Thirumalai, D. (2016). Salt Effects on the Thermodynamics of a Frameshifting RNA Pseudoknot under Tension. J. Mol. Biol. *428*, 2847–2859.
- Horiguchi, G., Mollá-Morales, A., Pérez-Pérez, J.M., Kojima, K., Robles, P., Ponce, M.R., Micol, J.L., and Tsukaya, H. (2011). Differential contributions of ribosomal protein genes to Arabidopsis thaliana leaf development. Plant J. *65*, 724–736.
- Huang, B., and Xu, C. (2008). Identification and characterization of proteins associated with plant tolerance to heat stress. J. Integr. Plant Biol. *50*, 1230–1237.
- Huang, I.C., Bosch, B.J., Li, F., Li, W., Kyoung, H.L., Ghiran, S., Vasilieva, N., Dermody, T.S., Harrison, S.C., Dormitzer, P.R., et al. (2006). SARS coronavirus, but not human coronavirus NL63, utilizes cathepsin L to infect ACE2-expressing cells. J. Biol. Chem. *281*, 3198–3203.
- Huang, W.M., Ao, S.Z., Casjens, S., Orlandi, R., Zeikus, R., Weiss, R., Winge, D., and Fang, M. (1988). A persistent untranslated sequence within bacteriophage T4 DNA topoisomerase gene 60. Science. *239*, 1005–1012.
- Huminiecki, L., and Wolfe, K.H. (2004). Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res. *14*, 1870–1879.
- Hummel, M., Cordewener, J.H.G., de Groot, J.C.M., Smeekens, S., America, A.H.P., and Hanson, J. (2012). Dynamic protein composition of Arabidopsis thaliana cytosolic ribosomes in response to sucrose feeding as revealed by label free MS E proteomics. Proteomics *12*, 1024–1038.
- Hurt, J.A., Robertson, A.D., and Burge, C.B. (2013). Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. Genome Res *23*, 1636–1650.
- Hussain, M., Torres, S., Schnettler, E., Funk, A., Grundhoff, A., Pijlman, G.P., Khromykh, A.A., and Asgari, S. (2012). West Nile virus encodes a microRNA-like small RNA in the 3' untranslated region which up-regulates GATA4 mRNA and facilitates virus replication in mosquito cells. Nucleic Acids Res. *40*, 2210–2223.
- Hussmann, J.A., Patchett, S., Johnson, A., Sawyer, S., and Press, W.H. (2015). Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. PLoS Genet. *11*, e1005732.
- Hütter, G., Nowak, D., Mossner, M., Ganepola, S., Müssig, A., Allers, K., Schneider, T., Hofmann, J., Kücherer, C., Blau, O., et al. (2009). Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. N. Engl. J. Med. *360*, 692–698.
- Imai, A., Hanzawa, Y., Komura, M., Yamamoto, K.T., Komeda, Y., and Takahashi, T. (2006). The dwarf phenotype of the Arabidopsis acl5 mutant is suppressed by a mutation in an upstream ORF of a bHLH gene. Development *133*, 3575–3585.
- Immanuel, T.M., Greenwood, D.R., and MacDiarmid, R.M. (2012). A critical review of translation initiation factor eIF2 kinases in plants - Regulating protein synthesis during stress. Funct. Plant Biol. *39*, 717–735.
- Ingolia, N.T. (2016). Ribosome Footprint Profiling of Translation throughout the Genome. Cell *165*, 22–33.
- Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-Wide Analysis in Vivo of Resolution Using Ribosome Profiling. Science. *324*, 218–223.
- Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802.

- Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., and Weissman, J.S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc *7*, 1534–1550.
- Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J.S., Jackson, S.E., Wills, M.R., and Weissman, J.S. (2014). Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. Cell Rep. *8*, 1365–1379.
- Ireland, D.D.C., Stohlman, S. a, Hinton, D.R., Atkinson, R., and Bergmann, C.C. (2008). Type I interferons are essential in controlling neurotropic coronavirus infection irrespective of functional CD8 T cells. J. Virol. *82*, 300–310.
- Irigoyen, N., Firth, A.E., Jones, J.D., Chung, B.Y., Siddell, S.G., and Brierley, I. (2016). High-Resolution Analysis of Coronavirus Gene Expression by RNA Sequencing and Ribosome Profiling. PLoS Pathog *12*, e1005473.
- Issur, M., Geiss, B.J., Bougie, I., Picard-Jean, F., Despins, S., Mayette, J., Hobdey, S.E., and Bisaillon, M. (2009). The flavivirus NS5 protein is a true RNA guanylyltransferase that catalyzes a two-step reaction to form the RNA cap structure. RNA *15*, 2340–2350.
- Ito, K., and Chiba, S. (2013). Arrest peptides: cis-acting modulators of translation. Annu. Rev. Biochem. *82*, 171–202.
- Ivanov, I.P., Firth, A.E., Michel, A.M., Atkins, J.F., and Baranov, P. V. (2011). Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Res. *39*, 4220–4234.
- Iwakura, N., Yokoyama, T., Quaglia, F., Mitsuoka, K., Mio, K., Shigematsu, H., Shirouzu, M., Kaji, A., and Kaji, H. (2017). Chemical and structural characterization of a model Post-Termination Complex (PoTC) for the ribosome recycling reaction: Evidence for the release of the mRNA by RRF and EF-G. PLoS One *12*, e0177972.
- Jacks, T., and Varmus, H.E. (1985). Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting. Science *230*, 1237–1242.
- Jackson, R., and Standart, N. (2015). The awesome power of ribosome profiling. RNA *21*, 652–654.
- Jackson, R.J., Hellen, C.U.T., and Pestova, T. V (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. Nat. Rev. Mol. Cell Biol. *11*, 113–127.
- Jackson, R.J., Hellen, C.U.T., and Pestova, T. V. (2012). Termination and post-termination events in eukaryotic translation. Adv. Protein Chem. Struct. Biol. *86*, 45–93.
- Jan, C.H., Williams, C.C., and Weissman, J.S. (2014). Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. Science. *346*, 1257521.
- Jang, S.K., Kräusslich, H.G., Nicklin, M.J., Duke, G.M., Palmenberg, A.C., and Wimmer, E. (1988). A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. J Virol *62*, 2636–2643.
- Janosi, L., Shimizu, I., and Kaji, A. (1994). Ribosome recycling factor (ribosome releasing factor) is essential for bacterial growth. Proc. Natl. Acad. Sci. U. S. A. *91*, 4249–4253.
- Janssen, S., and Giegerich, R. (2014). The RNA shapes studio. Bioinformatics *31*, 423–425.
- Jensen, B.C., Ramasamy, G., Vasconcelos, E.J.R., Ingolia, N.T., Myler, P.J., and Parsons, M. (2014). Extensive stage-regulation of translation revealed by ribosome profiling of Trypanosoma brucei. BMC Genomics *15*, 911.
- Ji, Z., Song, R., Huang, H., Regev, A., and Struhl, K. (2016). Transcriptome-scale RNase-footprinting of RNA-protein complexes. Nat. Biotechnol. *34*, 410–413.

- Jia, Y., Polunovsky, V., Bitterman, P.B., and Wagner, C.R. (2012). Cap-Dependent Translation Initiation Factor eIF4E: An Emerging Anticancer Drug Target. Med. Res. Rev. *32*, 786–814.
- Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. Genome Res. *21*, 1543–1551.
- Jiang, X.-S., Tang, L.-Y., Dai, J., Zhou, H., Li, S.-J., Xia, Q.-C., Wu, J.-R., and Zeng, R. (2005). Quantitative analysis of severe acute respiratory syndrome (SARS)-associated coronavirus-infected cells using proteomic approaches: implications for cellular responses to virus infection. Mol. Cell. Proteomics *4*, 902–913.
- Jindadamrongwech, S., Thepparit, C., and Smith, D.R. (2004). Identification of GRP 78 (BiP) as a liver cell expressed receptor element for dengue virus serotype 2. Arch. Virol. *149*, 915–927.
- Johansson, M., Chen, J., Tsai, A., Kornberg, G., and Puglisi, J. (2014). Sequence-dependent elongation dynamics on macrolide-bound ribosomes. Cell Rep. *7*, 1534–1546.
- Jorgensen, R. a., and Dorantes-Acosta, A.E. (2012). Conserved Peptide Upstream Open Reading Frames are Associated with Regulatory Genes in Angiosperms. Front. Plant Sci. *3*, 1–11.
- Jousse, C., Bruhat, A., Carraro, V., Urano, F., Ferrara, M., Ron, D., and Fafournoux, P. (2001). Inhibition of CHOP translation by a peptide encoded by an open reading frame localized in the chop 5'UTR. Nucleic Acids Res. *29*, 4341–4351.
- Juntawong, P., and Bailey-Serres, J. (2012). Dynamic Light Regulation of Translation Status in Arabidopsis thaliana. Front. Plant Sci. *3*, 1–16.
- Juntawong, P., Sorenson, R., and Bailey-Serres, J. (2013). Cold shock protein 1 chaperones mRNAs during translation in Arabidopsis thaliana. Plant J. *74*, 1016–1028.
- Juntawong, P., Girke, T., Bazin, J., and Bailey-Serres, J. (2014). Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. Proc. Natl Acad. Sci. USA *111*, E203–E212.
- Kahles, A., Behr, J., and Rätsch, G. (2015). MMR: A tool for read multi-mapper resolution. Bioinformatics *32*, 770–772.
- Kannan, K. (2014). The general mode of translation inhibition by macrolide antibiotics. Proc. Natl Acad. Sci. USA *111*, 15958–15963.
- Kannan, K., Vazquez-Laslop, N., and Mankin, A.S. (2012). Selective protein synthesis by ribosomes with a drug-obstructed exit tunnel. Cell *151*, 508–520.
- Kapp, L.D., and Lorsch, J.R. (2004). The molecular mechanics of eukaryotic translation. Annu. Rev. Biochem. *73*, 657–704.
- Karijolich, J., and Yu, Y.T. (2014). Therapeutic suppression of premature termination codons: Mechanisms and clinical considerations (Review). Int. J. Mol. Med. *34*, 355–362.
- Karn, J., and Stoltzfus, C.M. (2012). Transcriptional and posttranscriptional regulation of HIV-1 gene expression. Cold Spring Harb Perspect Med *2*, a006916.
- Katz, Y., Li, F., Lambert, N.J., Sokol, E.S., Tam, W.L., Cheng, A.W., Airoldi, E.M., Lengner, C.J., Gupta, P.B., Yu, Z., et al. (2014). Musashi proteins are post-transcriptional regulators of the epithelial-luminal cell state. Elife *3*, e03915.
- Kawaguchi, R., and Bailey-Serres, J. (2005). mRNA sequence features that contribute to translational regulation in Arabidopsis. Nucleic Acids Res. *33*, 955–965.

- Keeling, K.M., Xue, X., Gunn, G., and Bedwell, D.M. (2014). Therapeutics Based on Stop Codon Readthrough. Annu. Rev. Genomics Hum. Genet. 1–24.
- Kelly, A.E., and Goulden, M.L. (2008). Rapid shifts in plant distribution with recent climate change. Proc. Natl. Acad. Sci. U. S. A. *105*, 11823–11826.
- Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. Nature *467*, 103–107.
- Keshet, Y., and Seger, R. (2010). The MAP Kinase Signaling Cascades: A System of Hundreds of Components Regulates a Diverse Array of Physiological Functions in MAP Kinase Signaling Protocols. In MAP Kinase Signaling Protocols, pp. 3–38.
- Ketteler, R. (2012). On programmed ribosomal frameshifting: the alternative proteomes. Front. Genet. *3*, 242.
- Kharroubi, I., Ladrière, L., Cardozo, A.K., Dogusan, Z., Cnop, M., and Eizirik, D.L. (2004). Free fatty acids and cytokines induce pancreatic beta-cell apoptosis by different mechanisms: role of nuclear factor-kappaB and endoplasmic reticulum stress. Endocrinology *145*, 5087–5096.
- Khatter, H., Myasnikov, A.G., Natchiar, S.K., and Klaholz, B.P. (2015). Structure of the human 80S ribosome. Nature *520*, 640–645.
- Kieft, J.S., Rabe, J.L., and Chapman, E.G. (2015). New hypotheses derived from the structure of a flaviviral Xrn1-resistant RNA: Conservation, folding, and host adaptation. RNA Biol. *12*, 1169–1177.
- Kim, M., Pinto, S., Getnet, D., Nirujogi, R., Manda, S., and Chaerkady, R. (2014). A draft map of the human proteome. Nature *509*, 575–581.
- Kincaid, R.P., Burke, J.M., and Sullivan, C.S. (2012). RNA virus microRNA that mimics a B-cell oncomiR. Proc. Natl. Acad. Sci. U. S. A. *109*, 3077–3082.
- Kinney, R.M., Huang, C.Y.-H., Rose, B.C., Kroeker, A.D., Dreher, T.W., Iversen, P.L., and Stein, D.A. (2005). Inhibition of dengue virus serotypes 1 to 4 in vero cell cultures with morpholino oligomers. J. Virol. *79*, 5116–5128.
- Kirchner, S., and Ignatova, Z. (2014). Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. Nat. Rev. Genet. *16*, 98–112.
- Klase, Z.A., Sampey, G.C., and Kashanchi, F. (2013). Retrovirus infected cells contain viral microRNAs. Retrovirology *10*, 15.
- Klinge, S., Voigts-Hoffmann, F., Leibundgut, M., Arpagaus, S., and Ban, N. (2011). Crystal Structure of the Eukaryotic 60S Ribosomal Subunit in Complex with Initiation Factor 6. Science. *334*, 941–948.
- Kobayashi, Y., Zhuang, J., Peltz, S., and Dougherty, J. (2010). Identification of a cellular factor that modulates HIV-1 programmed ribosomal frameshifting. J. Biol. Chem. *285*, 19776–19784.
- Kochetov, A. V., Ahmad, S., Ivanisenko, V., Volkova, O.A., Kolchanov, N.A., and Sarai, A. (2008). uORFs, reinitiation and alternative translation start sites in human mRNAs. FEBS Lett. *582*, 1293–1297.
- Koetzner, C. a, Kuo, L., Goebel, S.J., Dean, A.B., Parker, M.M., and Masters, P.S. (2010). Accessory protein 5a is a major antagonist of the antiviral action of interferon against murine coronavirus. J. Virol. *84*, 8262–8274.
- Kondo, T., Hashimoto, Y., Kato, K., Inagaki, S., Hayashi, S., and Kageyama, Y. (2007). Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nat. Cell Biol. *9*, 660–665.

- Kozak, M. (1991). Structural features in eukaryotic mRNAs that modulate the initiation of translation. J. Biol. Chem. *266*, 19867–19870.
- Kramer, G., Boehringer, D., Ban, N., Bukau, B., Kramer, G., Boehringer, D., Ban, N., and Bukau, B. (2009). The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. Nat. Struct. & Mol. Biol. *16*, 589–597.
- Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J. Mol. Biol. *305*, 567–580.
- Ksiazek, T.G., Erdman, D., Goldsmith, C.S., Zaki, S.R., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J.A., Lim, W., et al. (2003). A novel coronavirus associated with severe acute respiratory syndrome. N. Engl. J. Med. *348*, 1953–1966.
- Kudelko, M., Brault, J.B., Kwok, K., Li, M.Y., Pardigon, N., Peiris, J.S.M., Bruzzone, R., Despre, P., Nal, B., and Wang, P.G. (2012). Class II ADP-ribosylation factors are required for efficient secretion of dengue viruses. J. Biol. Chem. *287*, 767–777.
- Kuersten, S., Radek, A., Vogel, C., and Penalva, L.O.F. (2013). Translation regulation gets its “omics” moment. Wiley Interdiscip. Rev. RNA *4*, 617–630.
- Kumar, S.V., and Wigge, P.A. (2010). H2A.Z-Containing Nucleosomes Mediate the Thermosensory Response in Arabidopsis. Cell *140*, 136–147.
- Kuo, L., Godeke, G.J., Raamsman, M.J., Masters, P.S., and Rottier, P.J. (2000). Retargeting of coronavirus by substitution of the spike glycoprotein ectodomain: crossing the host cell species barrier. J. Virol. *74*, 1393–1406.
- Kvam, V.M., Liu, P., and Si, Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am. J. Bot. *99*, 248–256.
- Lakkaraju, A.K.K., Mary, C., Scherrer, A., Johnson, A.E., and Strub, K. (2008). SRP Keeps Polypeptides Translocation-Competent by Slowing Translation to Match Limiting ER-Targeting Sites. Cell *133*, 440–451.
- Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., et al. (2012). The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. Nucleic Acids Res. *40*.
- Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Curr. Protoc. Bioinforma. *Chapter 11*, Unit 11.7.
- Larance, M., and Lamond, A.I. (2015). Multidimensional proteomics for cell biology. Nat. Rev. Mol. Cell Biol. *16*, 269–280.
- Lareau, L.F., Hite, D.H., Hogan, G.J., and Brown, P.O. (2014). Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. Elife *3*, e01257.
- Larsson, O., Sonenberg, N., and Nadon, R. (2011). anota: analysis of differential translation in genome-wide studies. Bioinforma. *27*, 1440–1441.
- Lau, S.K.P., and Chan, J.F.W. (2015). Coronaviruses: emerging and re-emerging pathogens in humans and animals. Virol. J. *12*, 209.
- Lazarowitz, S.G., and Robertson, H.D. (1977). Initiator regions from the small size class of reovirus messenger RNA protected by rabbit reticulocyte ribosomes. J. Biol. Chem. *252*, 7842–7849.
- Lee, A.S., Burdeinick-Kerr, R., and Whelan, S.P.J. (2013). A ribosome-specialized translation initiation pathway is required for cap-dependent translation of vesicular stomatitis virus mRNAs. Proc. Natl. Acad. Sci. U. S. A. *110*, 324–329.

- Lee, B., Henderson, D.A., and Zhu, J.-K. (2005). The Arabidopsis Cold-Responsive Transcriptome and Its Regulation by ICE1. Plant Cell Online *17*, 3155–3175.
- Lee, D.G., Ahsan, N., Lee, S.H., Kyu, Y.K., Jeong, D.B., Lee, I.J., and Lee, B.H. (2007). A proteomic approach in analyzing heat-responsive proteins in rice leaves. Proteomics *7*, 3369–3383.
- Lee, M.V., Topper, S.E., Hubler, S.L., Hose, J., Wenger, C.D., Coon, J.J., and Gasch, A.P. (2011). A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. Mol. Syst. Biol. *7*, 514.
- Lee, S., Weon, S., Lee, S., and Kang, C. (2010). Relative codon adaptation index, a sensitive measure of codon usage bias. Evol. Bioinforma. *2010*, 47–55.
- Lee, S., Liu, B., Lee, S., Huang, S.X., Shen, B., and Qian, S.B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A *109*, E2424-32.
- Lee, Y.Y., Cevallos, R.C., and Jan, E. (2009). An upstream open reading frame regulates translation of GADD34 during cellular stresses that induce eIF2phosphorylation. J. Biol. Chem. *284*, 6661–6673.
- Lei, L., Shi, J., Chen, J., Zhang, M., Sun, S., Xie, S., Li, X., Zeng, B., Peng, L., Hauck, A., et al. (2015). Ribosome profiling reveals dynamic translational landscape in maize seedlings under drought stress. Plant J. *84*, 1206–1208.
- Lellis, A.D., Allen, M.L., Aertker, A.W., Tran, J.K., Hillis, D.M., Harbin, C.R., Caldwell, C., Gallie, D.R., and Browning, K.S. (2010). Deletion of the eIFiso4G subunit of the Arabidopsis eIFiso4F translation initiation complex impairs health and viability. Plant Mol. Biol. *74*, 249–263.
- Lenoir, J., Gégout, J.C., Marquet, P.A., de Ruffray, P., and Brisse, H. (2008). A significant upward shift in plant species optimum elevation during the 20th century. Science *320*, 1768–1771.
- Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2009a). RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics *26*, 493–500.
- Li, G.-W.-. W., Oh, E., and Weissman, J.S. (2012a). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature *484*, 538–541.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009b). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
- Li, J., Witten, D.M., Johnstone, I.M., and Tibshirani, R. (2012b). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics *13*, 523–538.
- Li, Y., Humbert, S., and Howell, S.H. (2012c). ZmbZIP60 mRNA is spliced in maize in response to ER stress. BMC Res. Notes *5*, 144.
- Liang, X., Shen, W., Sun, H., Migawa, M.T., Vickers, T.A., and Crooke, S.T. (2016). Translation efficiency of mRNAs is increased by antisense oligonucleotides targeting upstream open reading frames. Nat. Biotechnol. *34*, 875–880.
- Lin, A.J., Zairis, S., Rabadan, R., Anzalone, A. V, and Cornish, V.W. (2016). Reprogramming eukaryotic translation with ligand-responsive synthetic RNA switches. Nat. Methods *13*, 1–8.
- Lin, Z., Gilbert, R.J.C., and Brierley, I. (2012). Spacer-length dependence of programmed-1 or-2 ribosomal frameshifting on a U6A heptamer supports a role for messenger RNA (mRNA) tension in frameshifting. Nucleic Acids Res. *40*, 8674–8689.

- Lindenbach, B.C., Murray, C.L., Heinz-Jurgen, T., and Rice, C.M. (2013). Flaviviridae. In Fields Virology. 6th Edition. II, H.P. Knipe DM, ed. (Wolters Kluwer, Lippincott Williams & Wilkens), pp. 712–745.
- Lissarre, M., Ohta, M., Sato, A., and Miura, K. (2010). Cold-responsive gene regulation during cold acclimation in plants. Plant Signal. Behav. *5*, 948–952.
- Lissenberg, A., Vrolijk, M.M., van Vliet, A.L.W., Langereis, M.A., de Groot-Mijnes, J.D.F., Rottier, P.J.M., and de Groot, R.J. (2005). Luxury at a cost? Recombinant mouse hepatitis viruses expressing the accessory hemagglutinin esterase protein display reduced fitness in vitro. J. Virol. *79*, 15054–15063.
- Liu, B., and Qian, S.B. (2014). Translational reprogramming in cellular stress response. Wiley Interdiscip. Rev. RNA *5*, 301–305.
- Liu, B., Han, Y., and Qian, S.B. (2013a). Cotranslational Response to Proteotoxic Stress by Elongation Pausing of Ribosomes. Mol. Cell *49*, 453–463.
- Liu, M.-J., Wu, S.-H.S.-H., Wu, J.-F., Lin, W.-D., Wu, Y.-C., Tsai, T.-Y., Tsai, H.-L., and Wu, S.-H.S.-H. (2013b). Translational landscape of photomorphogenic Arabidopsis. Plant Cell *25*, 3699–3710.
- Liu, Y., Zhou, J., and White, K.P. (2014). RNA-seq differential expression studies: More sequence or more replication? Bioinformatics *30*, 301–304.
- Lodish, H., Berk, A., and Zipursky, S. (2000). Processing of rRNA and tRNA. In Molecular Cell Biology, p. Section 11.6.
- Lopinski, J.D., Dinman, J.D., and Bruenn, J.A. (2000). Kinetics of Ribosomal Pausing during Programmed -1 Translational Frameshifting. Mol. Cell. Biol. *20*, 1095–1103.
- Loughran, G., Chou, M.Y., Ivanov, I.P., Jungreis, I., Kellis, M., Kiran, A.M., Baranov, P. V., and Atkins, J.F. (2014). Evidence of efficient stop codon readthrough in four mammalian genes. Nucleic Acids Res. *42*, 8928–8938.
- Lovén, J., Orlando, D.A., Sigova, A.A., Lin, C.Y., Rahl, P.B., Burge, C.B., Levens, D.L., Lee, T.I., and Young, R.A. (2012). Revisiting global gene expression analysis. Cell *151*, 476–482.
- Lozach, P.Y., Burleigh, L., Staropoli, I., Navarro-Sanchez, E., Harriague, J., Virelizier, J.L., Rey, F.A., Desprès, P., Arenzana-Seisdedos, F., and Amara, A. (2005). Dendritic cell-specific intercellular adhesion molecule 3-grabbing non-integrin (DC-SIGN)-mediated enhancement of dengue virus infection is independent of DC-SIGN internalization signals. J. Biol. Chem. *280*, 23698–23708.
- Lu, P.D., Harding, H.P., and Ron, D. (2004). Translation reinitiation at alternative open reading frames regulates gene expression in an integrated stress response. J. Cell Biol. *167*, 27–33.
- Lyubimova, A., Itzkovitz, S., Junker, J.P., Fan, Z.P., Wu, X., and van Oudenaarden, A. (2013). Single-molecule mRNA detection and counting in mammalian tissue. Nat. Protoc. *8*, 1743–1758.
- Ma, J. (2014). Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. J. Proteome Res. *13*, 1757–1765.
- Ma, J., Haldar, S., Khan, M. a, Sharma, S. Das, Merrick, W.C., Theil, E.C., and Goss, D.J. (2012). Fe2+ binds iron responsive element-RNA, selectively changing protein-binding affinities and regulating mRNA repression and activation. Proc. Natl. Acad. Sci. U. S. A. *109*, 8417–8422.
- Maldonado-Bonilla, L.D. (2014). Composition and function of P bodies in Arabidopsis thaliana. Front. Plant Sci. *5*, 201.

- Malik, V., Rodino-Klapac, L.R., Viollet, L., and Mendell, J.R. (2010). Aminoglycoside-induced mutation suppression (stop codon readthrough) as a therapeutic strategy for Duchenne muscular dystrophy. Ther. Adv. Neurol. Disord. *3*, 379–389.
- Manktelow, E., Shigemoto, K., and Brierley, I. (2005). Characterization of the frameshift signal of Edr, a mammalian example of programmed -1 ribosomal frameshifting. Nucleic Acids Res. *33*, 1553–1563.
- Manuell, A.L., Quispe, J., and Mayfield, S.P. (2007). Structure of the chloroplast ribosome: Novel domains for translation regulation. PLoS Biol. *5*, 1785–1797.
- Marceau, C.D., Puschnik, A.S., Majzoub, K., Ooi, Y.S., Brewer, S.M., Fuchs, G., Swaminathan, K., Mata, M.A., Elias, J.E., Sarnow, P., et al. (2016). Genetic dissection of Flaviviridae host factors through genome-scale CRISPR screens. Nature *535*, 1–20.
- Marcheschi, R.J., Mouzakis, K.D., and Butcher, S.E. (2009). Selection and characterization of small molecules that bind the HIV-1 frameshift site RNA. ACS Chem. Biol. *4*, 844–854.
- Marcheschi, R.J., Tonelli, M., Kumar, A., and Butcher, S.E. (2011). Structure of the HIV-1 frameshift site RNA bound to a small molecule inhibitor of viral replication. ACS Chem. Biol. *6*, 857–864.
- Marciniak, S.J., Yun, C.Y., Oyadomari, S., Novoa, I., Zhang, Y., Jungreis, R., Nagata, K., Harding, H.P., and Ron, D. (2004). CHOP induces death by promoting protein synthesis and oxidation in the stressed endoplasmic reticulum. Genes Dev. *18*, 3066–3077.
- Marsaglia, G., Tsang, W.W., and Wang, J. (2003). Evaluating Kolmogorov's Distribution. J. Stat. Softw. *8*, 1–4.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. *17*, 10.
- Mason, P.W. (1989). Maturation of Japanese encephalitis virus glycoproteins produced by infected mammalian and mosquito cells. Virology *169*, 354–364.
- Masters, P.S. (2006). The Molecular Biology of Coronaviruses. Adv. Virus Res. *65*, 193–292.
- Masters, P.S., and Perlman, S. (2013). Coronaviridae. In Fields Virology. 6th Edition. II, H.P. Knipe DM, ed. (Wolters Kluwer, Lippincott Williams & Wilkens), pp. 826–858.
- Matsufuji, S., Matsufuji, T., Miyazaki, Y., Murakami, Y., Atkins, J.F., Gesteland, R.F., and Hayashi, S. ichi (1995). Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme. Cell *80*, 51–60.
- Matsuura, H., Shinmyo, A., and Kato, K. (2008). Preferential translation mediated by Hsp81-3 5'-UTR during heat shock involves ribosome entry at the 5'-end rather than an internal site in Arabidopsis suspension cells. J. Biosci. Bioeng. *105*, 39–47.
- Matsuura, H., Ishibashi, Y., Shinmyo, A., Kanaya, S., and Kato, K. (2010). Genome-wide analyses of early translational responses to elevated temperature and high salinity in arabidopsis thaliana. Plant Cell Physiol. *51*, 448–462.
- Matsuura, H., Takenami, S., Kubo, Y., Ueda, K., Ueda, A., Yamaguchi, M., Hirata, K., Demura, T., Kanaya, S., and Kato, K. (2013). A computational and experimental approach reveals that the 5'-proximal region of the 5'-UTR has a Cis-regulatory signature responsible for heat stress-regulated mRNA translation in arabidopsis. Plant Cell Physiol. *54*, 474–483.
- Matusan, A.E., Pryor, M.J., Davidson, A.D., and Wright, P.J. (2001). Mutagenesis of the Dengue Virus Type 2 NS3 Protein within and outside Helicase Motifs : Effects on Enzyme Activity and Virus Replication. J. Virol. *75*, 9633–9643.

- Mayberry, L.K., Allen, M.L., Nitka, K.R., Campbell, L., Murphy, P.A., and Browning, K.S. (2011). Plant cap-binding complexes eukaryotic initiation factors eIF4F and eIFISO4F: Molecular specificity of subunit binding. J. Biol. Chem. *286*, 42566–42574.
- McGilvray, I.D., Lu, Z., Wei, A.C., Dackiw, A.P.B., Marshall, J.C., Kapus, A., Levy, G., and Rotstein, O.D. (1999). Murine hepatitis virus strain 3 induces the macrophage prothrombinase fgl-2 through p38 mitogen-activated protein kinase activation. J. Biol. Chem. *273*, 32222–32229.
- McLean, M.J., Wolfe, K.H., and Devine, K.M. (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J. Mol. Evol. *47*, 691–696.
- Melian, E.B., Hinzman, E., Nagasaki, T., Firth, A.E., Wills, N.M., Nouwens, A.S., Blitvich, B.J., Leung, J., Funk, A., Atkins, J.F., et al. (2010). NS1' of flaviviruses in the Japanese encephalitis virus serogroup is a product of ribosomal frameshifting and plays a role in viral neuroinvasiveness. J. Virol. *84*, 1641–1647.
- Melian, E.B., Hall-Mendelin, S., Du, F., Owens, N., Bosco-Lauth, A.M., Nagasaki, T., Rudd, S., Brault, A.C., Bowen, R.A., Hall, R.A., et al. (2014). Programmed Ribosomal Frameshift Alters Expression of West Nile Virus Genes and Facilitates Virus Replication in Birds and Mosquitoes. PLoS Pathog. *10*, e1004447.
- Melnikov, S., Ben-Shem, A., Garreau de Loubresse, N., Jenner, L., Yusupova, G., and Yusupov, M. (2012). One core, two shells: bacterial and eukaryotic ribosomes. Nat Struct Mol Biol *19*, 560–567.
- Menschaert, G., Van Criekinge, W., Notelaers, T., Koch, A., Crappé, J., Gevaert, K., and Van Damme, P. (2013). Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. Mol. Cell. Proteomics *12*, 1780–1790.
- Michael, T.P., Mockler, T.C., Breton, G., McEntee, C., Byer, A., Trout, J.D., Hazen, S.P., Shen, R., Priest, H.D., Sullivan, C.M., et al. (2008). Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. PLoS Genet. *4*, e0004921.
- Michel, A.M., Fox, G., M. Kiran, A., De Bo, C., O'Connor, P.B.F., Heaphy, S.M., Mullan, J.P.A., Donohue, C.A., Higgins, D.G., Baranov, P. V., et al. (2014). GWIPS-viz: Development of a ribo-seq genome browser. Nucleic Acids Res. *42*, D859–D864.
- Michel, A.M., Mullan, J.P.A.A., Velayudhan, V., O'Connor, P.B.F.F., Donohue, C.A., and Baranov, P. V. (2016). RiboGalaxy: a browser based platform for the alignment, analysis and visualization of ribosome profiling data. RNA Biol. *13*, 316–319.
- Miller, S., and Krijnse-Locker, J. (2008). Modification of intracellular membrane structures for virus replication. Nat. Rev. Microbiol. *6*, 363–374.
- Miller, J.L., deWet, B.J.M., Martinez-Pomares, L., Radcliffe, C.M., Dwek, R. a, Rudd, P.M., and Gordon, S. (2008). The Mannose Receptor Mediates Dengue Virus Infection of Macrophages. PLoS Pathog. *4*, 11.
- Miller, S., Kastner, S., Krijnse-Locker, J., Bühler, S., and Bartenschlager, R. (2007). The non-structural protein 4A of dengue virus is an integral membrane protein inducing membrane alterations in a 2K-regulated manner. J. Biol. Chem. *282*, 8873–8882.
- Mittler, R. (2006). Abiotic stress, the field environment and stress combination. Trends Plant Sci. *11*, 15–19.
- Mochizuki, K., Oguro, A., Ohtsu, T., Sonenberg, N., and Nakamura, Y. (2005). High affinity RNA for mammalian initiation factor 4E interferes with mRNA-cap binding and inhibits translation. RNA *11*, 77–89.

- Mohammad, F., Woolstenhulme, C.J., Green, R., and Buskirk, A.R. (2016). Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. Cell Rep. *14*, 686–694.
- Momattin, H., Mohammed, K., Zumla, A., Memish, Z. a, and Al-Tawfiq, J. a (2013). Therapeutic options for Middle East respiratory syndrome coronavirus (MERS-CoV)--possible lessons from a systematic review of SARS-CoV therapy. Int. J. Infect. Dis. *17*, e792-8.
- Moreno, A.A., and Orellana, A. (2011). The physiological role of the unfolded protein response in plants. Biol. Res. *44*, 75–80.
- Morisaki, T., Lyon, K., DeLuca, K.F., DeLuca, J.G., English, B.P., Zhang, Z., Lavis, L.D., Grimm, J.B., Viswanathan, S., Looger, L.L., et al. (2016). Real-time quantification of single RNA translation dynamics in living cells. Science *352*, 1425-1429.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.
- Mouilleron, H., Delcourt, V., and Roucou, X. (2016). Death of a dogma: Eukaryotic mRNAs can code for more than one protein. Nucleic Acids Res. *44*, 14–23.
- Mouzakis, K.D., Lang, A.L., Vander Meulen, K.A., Easterday, P.D., and Butcher, S.E. (2013). HIV-1 frameshift efficiency is primarily determined by the stability of base pairs positioned at the mRNA entrance channel of the ribosome. Nucleic Acids Res. *41*, 1901–1913.
- Muench, D.G., Zhang, C., and Dahodwala, M. (2012). Control of cytoplasmic translation in plants. Wiley Interdiscip. Rev. RNA *3*, 178–194.
- Mulvey, M., Arias, C., and Mohr, I. (2007). Maintenance of endoplasmic reticulum (ER) homeostasis in herpes simplex virus type 1-infected cells through the association of a viral glycoprotein with PERK, a cellular ER stress sensor. J. Virol. *81*, 3377–3390.
- Muñoz-Jordan, J.L., Sánchez-Burgos, G.G., Laurent-Rolle, M., and García-Sastre, A. (2003). Inhibition of interferon signaling by dengue virus. Proc. Natl. Acad. Sci. U. S. A. *100*, 14333–14338.
- Mustafa, M.S., Rasotgi, V., Jain, S., and Gupta, V. (2015). Discovery of fifth serotype of dengue virus (denv-5): A new public health dilemma in dengue control. Med. J. Armed Forces India *71*, 67–70.
- Mustroph, A., Zanetti, M.E., Jang, C.J.H., Holtan, H.E., Repetti, P.P., Galbraith, D.W., Girke, T., and Bailey-Serres, J. (2009). Profiling translatomes of discrete cell populations resolves altered cellular priorities during hypoxia in Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. *106*, 18843–18848.
- Nagarajan, V.K., Jones, C.I., Newbury, S.F., and Green, P.J. (2013). XRN 5'→3' exoribonucleases: Structure, mechanisms and functions. Biochim. Biophys. Acta - Gene Regul. Mech. *1829*, 590–603.
- Nakamura, Y., Gojobori, T., and Ikemura, T. (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000 [In Process Citation]. Nucleic Acids Res *28*, 292.
- Nakhoul, H., Ke, J., Zhou, X., Liao, W., Zeng, S.X., and Lu, H. (2014). Ribosomopathies: Mechanisms of disease. Clin. Med. Insights Blood Disord. *7*, 7–16.
- Namy, O., Moran, S.J., Stuart, D.I., Gilbert, R.J.C., and Brierley, I. (2006). A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. Nature *441*, 244–247.
- Napierala, M. (2012). What Is the Bonferroni Correction ? AAOS Now *April*, 1–3.

- Napthine, S., Treffers, E.E., Bell, S., Goodfellow, I., Fang, Y., Firth, A.E., Snijder, E.J., and Brierley, I. (2016). A novel role for poly(C) binding proteins in programmed ribosomal frameshifting. Nucleic Acids Res *44*, gkw480.
- Navarro-Sanchez, E., Altmeyer, R., Amara, A., Schwartz, O., Fieschi, F., Virelizier, J.L., Arenzana-Seisdedos, F., and Despres, P. (2003). Dendritic-cell-specific ICAM3-grabbing non-integrin is essential for the productive infection of human dendritic cells by mosquito-cell-derived dengue viruses. EMBO Rep *4*, 723–728.
- Neuhaus, K., Landstorfer, R., Fellner, L., Simon, S., Schafferhans, A., Goldberg, T., Marx, H., Ozoline, O.N., Rost, B., Kuster, B., et al. (2016). Translatomics combined with transcriptomics and proteomics reveals novel functional, recently evolved orphan genes in Escherichia coli O157:H7 (EHEC). BMC Genomics *17*, 133.
- Neuman, B.W., Joseph, J.S., Saikatendu, K.S., Serrano, P., Chatterjee, A., Johnson, M.A., Liao, L., Klaus, J.P., Yates, J.R., Wuthrich, K., et al. (2008). Proteomics Analysis Unravels the Functional Repertoire of Coronavirus Nonstructural Protein 3. J Virol *82*, 5279–5294.
- Nicolaï, M., Roncato, M. a, Canoy, a S., Rouquié, D., Sarda, X., Freyssinet, G., and Robaglia, C. (2006). Large-scale analysis of mRNA translation states during sucrose starvation in arabidopsis cells identifies cell proliferation and chromatin structure as targets of translational control. Plant Physiol. *141*, 663–673.
- Nilsson, O.B., Hedman, R., Marino, J., Wickles, S., Bischoff, L., Johansson, M., Müller-Lucks, A., Trovato, F., Puglisi, J.D., O'Brien, E.P., et al. (2015). Cotranslational Protein Folding inside the Ribosome Exit Tunnel. Cell Rep. *12*, 1533–1540.
- Nishimura, T., Wada, T., Yamamoto, K.T., and Okada, K. (2005). The Arabidopsis STV1 protein, responsible for translation reinitiation, is required for auxin-mediated gynoecium patterning. Plant Cell *17*, 2940–2953.
- Noble, W.S. (2009). How does multiple testing correction work? Nat. Biotechnol. *27*, 1135–1137.
- Noisakran, S., Dechtawewat, T., Avirutnan, P., Kinoshita, T., Siripanyaphinyo, U., Puttikhunt, C., Kasinrerk, W., Malasit, P., and Sittisombut, N. (2008). Association of dengue virus NS1 protein with lipid rafts. J. Gen. Virol. *89*, 2492–2500.
- Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., Uhlén, M., and Nielsen, J. (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: A case study in Saccharomyces cerevisiae. Nucleic Acids Res. *40*, 10084–10097.
- Normile, D. (2013). Surprising New Dengue Virus Throws A Spanner in Disease Control Efforts. Science. *342*, 2013.
- De Nova-Ocampo, M., Villegas-Sepúlveda, N., and del Angel, R.M. (2002). Translation elongation factor-1alpha, La, and PTB interact with the 3' untranslated region of dengue 4 virus RNA. Virology *295*, 337–347.
- Nürenberg, E., and Tampé, R. (2013). Tying up loose ends: Ribosome recycling in eukaryotes and archaea. Trends Biochem. Sci. *38*, 64–74.
- O'Connor, P.B.F., Li, G.W., Weissman, J.S., Atkins, J.F., and Baranov, P. V. (2013). RRNA:mRNA pairing alters the length and the symmetry of mRNA-protected fragments in ribosome profiling experiments. Bioinformatics *29*, 1488–1491.
- O'Connor, P.B.F., Andreev, D.E., and Baranov, P. V. (2016). Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. Nat. Commun. *7*, 12915.

- Obermayer, B., and Rajewsky, N. (2014). Mixed messages: Re-initiation factors regulate translation of animal mRNAs. Cell Res. *24*, 1383–1384.
- Olexiouk, V., Crappé, J., Verbruggen, S., Verhegen, K., Martens, L., and Menschaert, G. (2015). sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. *44*, D324-9.
- Olshen, A.B., Hsieh, A.C., Stumpf, C.R., Olshen, R.A., Ruggero, D., and Taylor, B.S. (2013). Assessing gene-level translational control from ribosome profiling. Bioinformatics *29*, 2995–3002.
- Oostra, M., te Lintelo, E.G., Deijs, M., Verheije, M.H., Rottier, P.J.M., and de Haan, C. a M. (2007). Localization and membrane topology of coronavirus nonstructural protein 4: involvement of the early secretory pathway in replication. J. Virol. *81*, 12323–12336.
- Ori, A., Toyama, B.H., Harris, M.S., Bock, T., Iskar, M., Bork, P., Ingolia, N.T., Hetzer, M.W., and Beck, M. (2015). Integrated Transcriptome and Proteome Analyses Reveal Organ-Specific Proteome Deterioration in Old Rats. Cell Syst. *1*, 224–237.
- Oshlack, A., and Wakefield, M. (2009). Transcript length bias in RNA-seq data confounds systems biology. Biol. Direct *4*, 14.
- Palam, L.R., Baird, T.D., and Wek, R.C. (2011). Phosphorylation of eIF2 facilitates ribosomal bypass of an inhibitory upstream ORF to enhance CHOP translation. J. Biol. Chem. *286*, 10939–10949.
- Paranjape, S.M., and Harris, E. (2007). Y box-binding protein-1 binds to the dengue virus 3'-untranslated region and mediates antiviral effects. J. Biol. Chem. *282*, 30497–30508.
- Paranjape, S.M., and Harris, E. (2009). Control of dengue virus translation and replication. Curr. Top. Microbiol. Immunol. *338*, 15–34.
- Park, N.I., and Muench, D.G. (2007). Biochemical and cellular characterization of the plant ortholog of PYM, a protein that interacts with the exon junction complex core proteins Mago and Y14. Planta *225*, 625–639.
- Park, S.-H., Chung, P.J., Juntawong, P., Bailey-Serres, J., Kim, Y.S., Jung, H., Bang, S.W., Kim, Y.-K., Do Choi, Y., and Kim, J.-K. (2012). Posttranscriptional Control of Photosynthetic mRNA Decay under Stress Conditions Requires 3' and 5' Untranslated Regions and Correlates with Differential Polysome Association in Rice. Plant Physiol. *159*, 1111–1124.
- Pasternak, A.O., Spaan, W.J.M., and Snijder, E.J. (2006). Nidovirus transcription: How to make sense...? J. Gen. Virol. *87*, 1403–1421.
- Pauli, A., Norris, M.L., Valen, E., Chew, G.-L., Gagnon, J. a, Zimmerman, S., Mitchell, A., Ma, J., Dubrulle, J., Reyon, D., et al. (2014). Toddler: an embryonic signal that promotes cell movement via Apelin receptors. Science. *343*, 746.
- Pavio, N., Romano, P.R., Graczyk, T.M., Feinstone, S.M., and Taylor, D.R. (2003). Protein synthesis and endoplasmic reticulum stress can be modulated by the hepatitis C virus envelope protein E2 through the eukaryotic initiation factor 2alpha kinase PERK. J. Virol. *77*, 3578–3585.
- Pechmann, S., Willmund, F., and Frydman, J. (2013). The Ribosome as a Hub for Protein Quality Control. Mol. Cell *49*, 411–421.
- Pelletier, J., and Sonenberg, N. (1988). Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. Nature *334*, 320–325.
- Pena, J., and Harris, E. (2011). Dengue virus modulates the unfolded protein response in a time-dependent manner. J. Biol. Chem. *286*, 14226–14236.

- Pena, J., and Harris, E. (2012). Early dengue virus protein synthesis induces extensive rearrangement of the endoplasmic reticulum independent of the UPR and SREBP-2 pathway. PLoS One *7*.
- Pendle, A.F., Clark, G.P., Boon, R., Lewandowska, D., Lam, Y.W., Andersen, J., Mann, M., Lamond, A.I., Brown, J.W., and Shaw, P.J. (2005). Proteomic analysis of the Arabidopsis nucleolus suggests novel nucleolar functions. Mol Biol Cell *16*, 260–269.
- Pérez Dominguez, M., Rojas, R., Bandeira, E., Requena, D., Ferreras, A.C., Triana, J.L., and Triana-Alonso, F. (2015). Traducción independiente de la estructura 5´cap del ARN genómico del virus dengue . Rev. Peru. Med. Exp. Y Salud Pública *32*, 11–18.
- Pesole, G., Liuni, S., Grillo, G., and Saccone, C. (1997). Structural and compositional features of untranslated regions of eukaryotic mRNAs. In Gene, pp. 95–102.
- Pesole, G., Grillo, G., Larizza, a, and Liuni, S. (2000). The untranslated regions of eukaryotic mRNAs: structure, function, evolution and bioinformatic tools for their analysis. Brief. Bioinform. *1*, 236–249.
- Pfeffer, S., Zavolan, M., Grässer, F.A., Chien, M., Russo, J.J., Ju, J., John, B., Enright, A.J., Marks, D., Sander, C., et al. (2004). Identification of Virus-Encoded MicroRNAs. Science. *304*, 734–736.
- Phillips, S.L., Soderblom, E.J., Bradrick, S.S., and Garcia-Blanco, M.A. (2016). Identification of proteins bound to dengue viral RNA In Vivo reveals new host proteins important for virus replication. MBio *7*.
- Phizicky, E.M., and Hopper, A.K. (2010). tRNA biology charges to the front. Genes Dev. *24*, 1832–1860.
- Piccirillo, C. a, Bjur, E., Topisirovic, I., Sonenberg, N., and Larsson, O. (2014). Translational control of immune responses: from transcripts to translatomes. Nat. Immunol. *15*, 503–511.
- Picotti, P., Bodenmiller, B., Mueller, L.N., Domon, B., and Aebersold, R. (2009). Full Dynamic Range Proteome Analysis of S. cerevisiae by Targeted Proteomics. Cell *138*, 795–806.
- Pierson, T.C., and Diamond, M.S. (2013). Flaviviruses. In Fields Virology. 6th Edition. II, H.P. Knipe DM, ed. (Wolters Kluwer, Lippincott Williams & Wilkens), pp. 747–793.
- Pijlman, G.P., Funk, A., Kondratieva, N., Leung, J., Torres, S., van der Aa, L., Liu, W.J., Palmenberg, A.C., Shi, P.Y., Hall, R.A., et al. (2008). A Highly Structured, Nuclease-Resistant, Noncoding RNA Produced by Flaviviruses Is Required for Pathogenicity. Cell Host Microbe *4*, 579–591.
- Pineau, L., Colas, J., Dupont, S., Beney, L., Fleurat-Lessard, P., Berjeaud, J.M., Bergs, T., and Ferreira, T. (2009). Lipid-induced ER stress: Synergistic effects of sterols and saturated fatty acids. Traffic *10*, 673–690.
- Piques, M., Schulze, W.X., Höhne, M., Usadel, B., Gibon, Y., Rohwer, J., and Stitt, M. (2009). Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in Arabidopsis. Mol. Syst. Biol. *5*, 314.
- Pisarev, A. V, Kolupaeva, V.G., Yusupov, M.M., Hellen, C.U.T., and Pestova, T. V (2008). Ribosomal position and contacts of mRNA in eukaryotic translation initiation complexes. EMBO J. *27*, 1609–1621.
- Plourde, A.R., and Bloch, E.M. (2016). A literature review of zika virus. Emerg. Infect. Dis. *22*, 1185–1192.

- Pöyry, T.A.A., Kaminski, A., and Jackson, R.J. (2004). What determines whether mammalian ribosomes resume scanning after translation of a short upstream open reading frame? Genes Dev. *18*, 62–75.
- Price, J.C., Guan, S., Burlingame, A., Prusiner, S.B., and Ghaemmaghami, S. (2010). Analysis of proteome dynamics in the mouse brain. Proc. Natl. Acad. Sci. U. S. A. *107*, 14508–14513.
- Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F.F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., et al. (2014). RefSeq: An update on mammalian reference sequences. Nucleic Acids Res. *42*, D756-63.
- Puckette, M., Iyer, N.J., Tang, Y., Dai, X. Bin, Zhao, P., and Mahalingam, R. (2012). Differential mRNA translation in medicago truncatula accessions with contrasting responses to ozone-induced oxidative stress. Mol. Plant *5*, 187–204.
- Puthalakath, H., O'Reilly, L.A., Gunn, P., Lee, L., Kelly, P.N., Huntington, N.D., Hughes, P.D., Michalak, E.M., McKimm-Breschkin, J., Motoyama, N., et al. (2007). ER Stress Triggers Apoptosis by Activating BH3-Only Protein Bim. Cell *129*, 1337–1349.
- Qian, W., Yang, J.-R.-. R., Pearson, N.M., Maclean, C., and Zhang, J. (2012). Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency. PLoS Genet *8*, e1002603.
- Qiu, Z., Hingley, S.T., Simmons, G., Yu, C., Das Sarma, J., Bates, P., and Weiss, S.R. (2006). Endosomal proteolysis by cathepsins is necessary for murine coronavirus mouse hepatitis virus type 2 spike-mediated entry. J. Virol. *80*, 5768–5776.
- Raaben, M., Groot Koerkamp, M.J.A., Rottier, P.J.M., and de Haan, C.A.M. (2007). Mouse hepatitis coronavirus replication induces host translational shutoff and mRNA decay, with concomitant formation of stress granules and processing bodies. Cell. Microbiol. *9*, 2218–2229.
- Raaben, M., Groot Koerkamp, M.J., Rottier, P.J., and de Haan, C.A. (2009). Type I interferon receptor-independent and -dependent host transcriptional responses to mouse hepatitis coronavirus infection in vivo. BMC Genomics *10*, 350.
- Ragin, A.D., Morgan, R.A., and Chmielewski, J. (2002). Cellular Import Mediated by Nuclear Localization Signal Peptide Sequences. Chem. Biol. *9*, 943–948.
- Rahmani, F., Hummel, M., Schuurmans, J., Wiese-Klinkenberg, A., Smeekens, S., and Hanson, J. (2009). Sucrose Control of Translation Mediated by an Upstream Open Reading Frame-Encoded Peptide. Plant Physiol. *150*, 1356–1367.
- Raj, A., Wang, S.H., Shim, H., Harpak, A., Li, Y.I., Engelmann, B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2016). Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. Elife *5*, e13328.
- Ramakrishnan, V. (2011). The Eukaryotic Ribosome. Science. *331*, 681-682.
- Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D., and Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol *14*, R95.
- Rasmussen, S., Barah, P., Suarez-Rodriguez, M.C., Bressendorff, S., Friis, P., Costantino, P., Bones, A.M., Nielsen, H.B., and Mundy, J. (2013). Transcriptome responses to combinations of stresses in Arabidopsis. Plant Physiol. *161*, 1783–1794.
- Rawlinson, S.M., Pryor, M.J., Wright, P.J., and Jans, D.A. (2009). CRM1-mediated nuclear export of dengue virus RNA polymerase NS5 modulates interleukin-8 induction and virus production. J. Biol. Chem. *284*, 15589–15597.

- Reeves, M.B., Davies, A.A., McSharry, B.P., Wilkinson, G.W., and Sinclair, J.H. (2007). Complex I binding by a virally encoded RNA regulates mitochondria-induced cell death. Science *316*, 1345–1348.
- Reggiori, F., Monastyrska, I., Verheije, M.H., Calì, T., Ulasli, M., Bianchi, S., Bernasconi, R., De Haan, C.A.M., and Molinari, M. (2010). Coronaviruses hijack the LC3-I-positive EDEMosomes, ER-derived vesicles exporting short-lived ERAD regulators, for replication. Cell Host Microbe *7*, 500–508.
- Reid, D.W., Chen, Q., Tay, A.S., Shenolikar, S., and Nicchitta, C. V (2014). The unfolded protein response triggers selective mRNA release from the endoplasmic reticulum. Cell *158*, 1362–1374.
- Reinhardt, J.A., and Jones, C.D. (2013). Two rapidly evolving genes contribute to male fitness in Drosophila. J. Mol. Evol. *77*, 246–259.
- Reyes-Del Valle, J., Chávez-Salinas, S., Medina, F., and Del Angel, R.M. (2005). Heat shock protein 90 and heat shock protein 70 are components of dengue virus receptor complex in human cells. J. Virol. *79*, 4557–4567.
- Ribeiro, D.M., Araújo, W.L., Fernie, A.R., Schippers, J.H.M., and Mueller-Roeber, B. (2012). Translatome and metabolome effects triggered by gibberellins during rosette growth in Arabidopsis. J. Exp. Bot. *63*, 2769–2786.
- Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.
- Robert, C., and Watson, M. (2015). Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome Biol. *16*, 177.
- Roberts, A., and Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. Nat. Methods *10*, 71–73.
- Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L., and Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. *12*, R22.
- Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.
- Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.
- Robles, J. a, Qureshi, S.E., Stephen, S.J., Wilson, S.R., Burden, C.J., Taylor, J.M., Robles, A., Qureshi, S.E., Stephen, S.J., Wilson, S.R., et al. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics *13*, 484.
- Roby, J.A., Pijlman, G.P., Wilusz, J., and Khromykh, A.A. (2014). Noncoding subgenomic flavivirus RNA: Multiple functions in west Nile virus pathogenesis and modulation of host responses. Viruses *6*, 404–427.
- Rodnina, M. V., and Wintermeyer, W. (2009). Recent mechanistic insights into eukaryotic ribosomes. Curr. Opin. Cell Biol. *21*, 435–443.
- Ron, D., and Walter, P. (2007). Signal integration in the endoplasmic reticulum unfolded protein response. Nat. Rev. Mol. Cell Biol. *8*, 519–529.
- Rooijers, K., Loayza-Puch, F., Nijtmans, L.G., and Agami, R. (2013). Ribosome profiling reveals features of normal and disease-associated mitochondrial translation. Nat. Commun. *4*, 2886.

- Rosado, A., Li, R., van de Ven, W., Hsu, E., and Raikhel, N. V (2012). Arabidopsis ribosomal proteins control developmental programs through translational regulation of auxin response factors. Proc. Natl. Acad. Sci. U. S. A. *109*, 19537–19544.
- Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., et al. (2015). The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. *43*, D670–D681.
- Roth-Cross, J.K., Bender, S.J., and Weiss, S.R. (2008). Murine coronavirus mouse hepatitis virus is recognized by MDA5 and induces type I interferon in brain macrophages/microglia. J. Virol. *82*, 9829–9838.
- Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J.S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature *505*, 701–705.
- Roy, B., and von Arnim, A.G. (2013). Translational Regulation of Cytoplasmic mRNAs. Arabidopsis Book *11*, e0165.
- Ruben, S., Perkins, A., Purcell, R., Joung, K., Sia, R., Burghoff, R., Haseltine, W.A., and Rosen, C.A. (1989). Structural and functional characterization of human immunodeficiency virus tat protein. J Virol *63*, 1–8.
- Ryabova, L.A., Pooggin, M.M., and Hohn, T. (2006). Translation reinitiation and leaky scanning in plant viruses. Virus Res. *119*, 52–62.
- Ryu, K.-Y., Maehr, R., Gilchrist, C.A., Long, M.A., Bouley, D.M., Mueller, B., Ploegh, H.L., and Kopito, R.R. (2007). The mouse polyubiquitin gene UbC is essential for fetal liver development, cell-cycle progression and stress tolerance. EMBO J. *26*, 2693–2706.
- Sabchareon, A., Wallace, D., Sirivichayakul, C., Limkittikul, K., Chanthavanich, P., Suvannadabba, S., Jiwariyavej, V., Dulyachai, W., Pengsaa, K., Wartel, T.A., et al. (2012). Protective efficacy of the recombinant, live-attenuated, CYD tetravalent dengue vaccine in Thai schoolchildren: A randomised, controlled phase 2b trial. Lancet *380*, 1559–1567.
- Saghatelian, A., and Couso, J.P. (2015). Discovery and characterization of smORF-encoded bioactive polypeptides. Nat. Chem. Biol. *11*, 909–916.
- Sampath, A., and Padmanabhan, R. (2009). Molecular targets for flavivirus drug discovery. Antiviral Res. *81*, 6–15.
- Samson, M., Libert, F., Doranz, B.J., Rucker, J., Liesnard, C., Farber, C.M., Saragosti, S., Lapoumeroulie, C., Cognaux, J., Forceille, C., et al. (1996). Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature *382*, 722–725.
- Sarafianos, S.G., Marchand, B., Das, K., Himmel, D.M., Parniak, M.A., Hughes, S.H., and Arnold, E. (2009). Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition. J. Mol. Biol. *385*, 693–713.
- Sasaki, K., and Imai, R. (2012). Pleiotropic Roles of Cold Shock Domain Proteins in Plants. Front. Plant Sci. *2*, 1–6.
- Sasikumar, A.N., Perez, W.B., and Kinzy, T.G. (2012). The many roles of the eukaryotic elongation factor 1 complex. Wiley Interdiscip. Rev. RNA *3*, 543–555.
- Satoh, R., Nakashima, K., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2002). ACTCAT, a novel cis-acting element for proline- and hypoosmolarity-responsive expression of the ProDH gene encoding proline dehydrogenase in Arabidopsis. Plant Physiol. *130*, 709–719.
- Sawicki, S.G., Sawicki, D.L., and Siddell, S.G. (2007). A contemporary view of coronavirus transcription. J. Virol. *81*, 20–29.

- van der Schaar, H.M., Rust, M.J., Chen, C., van der Ende-Metselaar, H., Wilschut, J., Zhuang, X., and Smit, J.M. (2008). Dissecting the cell entry pathway of dengue virus by single-particle tracking in living cells. PLoS Pathog. *4*, e1000244.
- Schmidt, R.L., and Simonović, M. (2012). Synthesis and decoding of selenocysteine and human health. Croat. Med. J. *53*, 535–550.
- Schopman, N.C.T., Willemsen, M., Liu, Y.P., Bradley, T., Van Kampen, A., Baas, F., Berkhout, B., and Haasnoot, J. (2012). Deep sequencing of virus-infected cells reveals HIV-encoded small RNAs. Nucleic Acids Res. *40*, 414–427.
- Schueren, F. (2014). Peroxisomal lactate dehydrogenase is generated by translational readthrough in mammals. Elife *3*, e03640.
- Schulz, M.H., Zerbino, D.R., Vingron, M., and Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics *28*, 1086–1092.
- Schurch, N.J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G.G., Owen-Hughes, T., et al. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 1–13.
- Schwanhäusser, B., Busse, D., and Li, N. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.
- Schwartz, L.M., Halloran, M.E., Durbin, A.P., and Longini, I.M. (2015). The dengue vaccine pipeline: Implications for the future of dengue control. Vaccine *33*, 3293–3298.
- Schwartz, S., Felber, B.K., Benko, D.M., Fenyö, E.M., and Pavlakis, G.N. (1990). Cloning and functional analysis of multiply spliced mRNA species of human immunodeficiency virus type 1. J. Virol. *64*, 2519–2529.
- Seqc/Maqc-Iii Consortium (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat. Biotechnol. *32*, 903–914.
- Seyednasrollah, F., Laiho, A., and Elo, L.L. (2013). Comparison of software packages for detecting differential expression in RNA-seq studies. Brief. Bioinform. *16*, 59–70.
- Seyednasrollah, F., Rantanen, K., Jaakkola, P., and Elo, L.L. (2015). ROTS: reproducible RNA-seq biomarker detector-prognostic markers for clear cell renal cell cancer. Nucleic Acids Res. *44*, e1-e8.
- Shalgi, R., Hurt, J.A., Krykbaeva, I., Taipale, M., Lindquist, S., and Burge, C.B. (2013). Widespread Regulation of Translation by Elongation Pausing in Heat Shock. Mol. Cell *49*, 439–452.
- Sharma, M.R., Koc, E.C., Datta, P.P., Booth, T.M., Spremulli, L.L., and Agrawal, R.K. (2003). Structure of the mammalian mitochondrial ribosome reveals an expanded functional role for its component proteins. Cell *115*, 97–108.
- Sharma, P., Yan, F., Doronina, V.A., Escuin-Ordinas, H., Ryan, M.D., and Brown, J.D. (2012). 2A peptides provide distinct solutions to driving stop-carry on translational recoding. Nucleic Acids Res. *40*, 3143–3151.
- Sharp, P.M., and Li, W.-H. (1987). The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. *15*, 3021–3030.
- Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F., and Sockett, R.E. (2005). Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. *33*, 1141–1153.

- Shehu-Xhilaga, M., Crowe, S.M., and Mak, J. (2001). Maintenance of the Gag/Gag-Pol ratio is important for human immunodeficiency virus type 1 RNA dimerization and viral infectivity. J Virol *75*, 1834–1841.
- Shi, X., Khade, P.K., Sanbonmatsu, K.Y., and Joseph, S. (2012). Functional role of the sarcin-ricin loop of the 23s rRNA in the elongation cycle of protein synthesis. J. Mol. Biol. *419*, 125–138.
- Shisler, J.L., Senkevich, T.G., Berry, M.J., and Moss, B. (1998). Ultraviolet-induced cell death blocked by a selenoprotein from a human dermatotropic poxvirus. Science *279*, 102–105.
- Sidrauski, C., McGeachy, A.M., Ingolia, N.T., and Walter, P. (2015). The small molecule ISRIB reverses the effects of eIF2 phosphorylation on translation and stress granule assembly. Elife *4*, e00498.
- Silvera, D., Formenti, S.C., and Schneider, R.J. (2010). Translational control in cancer. Nat Rev Cancer *10*, 254–266.
- Simonović, M., and Steitz, T.A. (2009). A structural view on the mechanism of the ribosome-catalyzed peptide bond formation. Biochim. Biophys. Acta - Gene Regul. Mech. *1789*, 612–623.
- Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. *15*, 121–132.
- Siwiak, M., and Zielenkiewicz, P. (2010). A comprehensive, quantitative, and genome-wide model of translation. PLoS Comput. Biol. *6*, 4.
- Skinner, S.O., Sepúlveda, L. a, Xu, H., and Golding, I. (2013). Measuring mRNA copy number in individual Escherichia coli cells using single-molecule fluorescent in situ hybridization. Nat. Protoc. *8*, 1100–1113.
- Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L., and Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat. Chem. Biol. *9*, 59–64.
- Sloan, E. a, Kearney, M.F., Gray, L.R., Anastos, K., Daar, E.S., Margolick, J., Maldarelli, F., Hammarskjold, M.-L., and Rekosh, D. (2013). Limited nucleotide changes in the Rev response element (RRE) during HIV-1 infection alter overall Rev-RRE activity and Rev multimerization. J. Virol. *87*, 11173–11186.
- Smith, E.C. (2013). Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. PLoS Pathog. *9*, e1003565.
- Smith, E.C., and Denison, M.R. (2013). Coronaviruses as DNA Wannabes: A New Model for the Regulation of RNA Virus Replication Fidelity. PLoS Pathog. *9*, e1003760.
- Smits, P., Smeitink, J., Van Den Heuvel, L., Smits, P., and Smeitink, J. (2010). Mitochondrial translation and beyond: Processes implicated in combined oxidative phosphorylation deficiencies (Hindawi).
- Sola, I., Moreno, J.L., Zuniga, S., Alonso, S., and Enjuanes, L. (2005). Role of nucleotides immediately flanking the transcription-regulating sequence core in coronavirus subgenomic mRNA synthesis. J Virol *79*, 2506–2516.
- Somers, J., Pöyry, T., and Willis, A.E. (2013). A perspective on mammalian upstream open reading frame function. Int. J. Biochem. Cell Biol. *45*, 1690–1700.
- Sonenberg, N., and Hinnebusch, A.G. (2009). Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. Cell *136*, 731–745.

- Soneson, C., and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics *14*, 91.
- Song, H.C., Seo, M.Y., Stadler, K., Yoo, B.J., Choo, Q.L., Coates, S.R., Uematsu, Y., Harada, T., Greer, C.E., Polo, J.M., et al. (2004). Synthesis and characterization of a native, oligomeric form of recombinant severe acute respiratory syndrome coronavirus spike glycoprotein. J Virol *78*, 10328–10335.
- Song, J., Irwin, J., and Dean, C. (2013). Remembering the prolonged cold of winter.
- Sorenson, R., and Bailey-Serres, J. (2014). Selective mRNA sequestration by OLIGOURIDYLATE-BINDING PROTEIN 1 contributes to translational control during hypoxia in Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. *111*, 2373–2378.
- Spiegel, M., Pichlmair, A., Martínez-Sobrido, L., Cros, J., García-Sastre, A., Haller, O., and Weber, F. (2005). Inhibition of Beta interferon induction by severe acute respiratory syndrome coronavirus suggests a two-step model for activation of interferon regulatory factor 3. J. Virol. *79*, 2079–2086.
- Spriggs, K.A., Bushell, M., and Willis, A.E. (2010). Translational regulation of gene expression during conditions of cell stress. Mol. Cell *40*, 228–237.
- Stadler, M., and Fire, A. (2013). Conserved translatome remodeling in nematode species executing a shared developmental transition. PLoS Genet. *9*, e1003739.
- Stadler, M., Artiles, K., Pak, J., and Fire, A. (2012). Contributions of mRNA abundance, ribosome loading, and post- or peri-translational effects to temporal repression of C. elegans heterochronic miRNA targets. Genome Res *22*, 2418–2426.
- Starck, S.R., Tsai, J.C., Chen, K., Shodiya, M., Wang, L., Yahiro, K., Martins-Green, M., Shastri, N., and Walter, P. (2016). Translation from the 5' untranslated region shapes the integrated stress response. Science. *351*, aad3867.
- Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., Akerman, M., Alioto, T., Ambrosini, G., Antonarakis, S.E., Behr, J., Bertone, P., et al. (2013). Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods *10*, 1177–1184.
- Steitz, J.A. (1969). Polypeptide Chain Initiation: Nucleotide Sequences of the Threee Ribosomal Binding Sites in Bacteriophage R17 RNA. Nature *224*, 957–964.
- Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V.T.K., Hein, M.Y., Huang, S.-X.X., Ma, M., Shen, B., Qian, S.-B.B., Hengel, H., et al. (2012). Decoding human cytomegalovirus. Science. *338*, 1088–1093.
- Stewarta, A.F.R., Richard, C.W., Suzowa, J., Stephan, D., Weremowicze, S., Morton, C.C., and Adra, C.N. (1996). Cloning of human RTEF-1, a transcriptional enhancer factor-1-related gene preferentially expressed in skeletal muscle: evidence for an ancient multigene family. Genomics *37*, 68–76.
- Stockman, L.J., Bellamy, R., and Garner, P. (2006). SARS: Systematic review of treatment effects. PLoS Med. *3*, 1525–1531.
- Stoltzfus, C.M., and Madsen, J.M. (2006). Role of viral splicing elements and cellular RNA binding proteins in regulation of HIV-1 alternative RNA splicing. Curr. HIV Res. *4*, 43–55.
- Su, X., Yu, Y., Zhong, Y., Giannopoulou, E.G., Hu, X., Liu, H., Cross, J.R., Rätsch, G., Rice, C.M., and Ivashkiv, L.B. (2015). Interferon-γ regulates cellular metabolism and mRNA translation to potentiate macrophage activation. Nat. Immunol. *16*, 838–849.
- Subtelny, A.O., Eichhorn, S.W., Chen, G.R., Sive, H., and Bartel, D.P. (2014). Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature *508*, 66–71.
- Sugihara, Y., Honda, H., Lida, T., Morinaga, T., Hino, S., Okajima, T., Matsuda, T., and Nadano, D. (2010). Proteomic analysis of rodent ribosomes revealed heterogeneity

including ribosomal proteins L10-like, L22-like 1, and L39-like. J. Proteome Res. *9*, 1351–1366.

- Sung, S., and Amasino, R.M. (2005). Remembering Winter: Toward a Molecular Understanding of Vernalization. Annu. Rev. Plant Biol *56*, 491–508.
- Sunkar, R., Li, Y.F., and Jagadeeswaran, G. (2012). Functions of microRNAs in plant stress responses. Trends Plant Sci. *17*, 196–203.
- Szegezdi, E., Logue, S.E., Gorman, A.M., and Samali, A. (2006). Mediators of endoplasmic reticulum stress-induced apoptosis. EMBO Rep. *7*, 880–885.
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447–D452.
- Tabas, I., and Ron, D. (2011). Integrating the mechanisms of apoptosis induced by endoplasmic reticulum stress. Nat. Cell Biol. *13*, 184–190.
- Tajima, S., Nukui, Y., Takasaki, T., and Kurane, I. (2007). Characterization of the variable region in the 3′ non-translated region of dengue type 1 virus. J. Gen. Virol. *88*, 2214–2222.
- Takamatsu, Y., Okamoto, K., Dinh, D.T., Yu, F., Hayasaka, D., Uchida, L., Nabeshima, T., Buerano, C.C., and Morita, K. (2014). NS1′ protein expression facilitates production of Japanese encephalitis virus in avian cells and embryonated chicken eggs. J. Gen. Virol. *95*, 373–383.
- Takyar, S., Hickerson, R.P., and Noller, H.F. (2005). mRNA helicase activity of the ribosome. Cell *120*, 49–58.
- Taniguchi, I., Mabuchi, N., and Ohno, M. (2014). HIV-1 Rev protein specifies the viral RNA export pathway by suppressing TAP/NXF1 recruitment. Nucleic Acids Res. *42*, 6645–6658.
- Tarazona, S., Tari, P.F., Turra, D., Pietro, A. Di, Nueda, M.J., Ferrer, A., and Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R / Bioc package ıa Jos e. Nucleic Acids Res. 1–15.
- Tassaneetrithep, B., Burgess, T.H., Granelli-Piperno, A., Trumpfheller, C., Finke, J., Sun, W., Eller, M.A., Pattanapanyasat, K., Sarasombath, S., Birx, D.L., et al. (2003). DC-SIGN (CD209) mediates dengue virus infection of human dendritic cells. J. Exp. Med. *197*, 823–829.
- Thalor, S.K., Berberich, T., Lee, S.S., Yang, S.H., Zhu, X., Imai, R., Takahashi, Y., and Kusano, T. (2012). Deregulation of sucrose-controlled translation of a bZIP-type transcription factor results in sucrose accumulation in leaves. PLoS One *7*, e33111.
- Theil, E.C. (2015). IRE mRNA riboregulators use metabolic iron (Fe^{2+}) to control mRNA activity and iron chemistry in animals. Metallomics *7*, 15–24.
- Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.
- Toyama, B.H., Savas, J.N., Park, S.K., Harris, M.S., Ingolia, N.T., Yates, J.R., and Hetzer, M.W. (2013). Identification of long-lived proteins reveals exceptional stability of essential cellular structures. Cell *154*, 971–982.
- Tran, M.K., Schultz, C.J., and Baumann, U. (2008). Conserved upstream open reading frames in higher plants. BMC Genomics *9*, 361.

- Trapnell, C., and Salzberg, S.L. (2009). How to map billions of short reads onto genomes. Nat. Biotechnol. *27*, 455–457.
- Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.
- Trapnell, C., Williams, B. a, Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.
- Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. *31*, 46–53.
- Tsirigotis, M., Zhang, M., Chiu, R.K., Wouters, B.G., and Gray, D.A. (2001). Sensitivity of Mammalian Cells Expressing Mutant Ubiquitin to Protein-damaging Agents. J. Biol. Chem. *276*, 46073–46078.
- Tsuchihashi, Z., and Kornberg, A. (1990). Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. Proc. Natl. Acad. Sci. *87*, 2516–2520.
- Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010a). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell *141*, 344–354.
- Tuller, T., Waldman, Y.Y., Kupiec, M., and Ruppin, E. (2010b). Translation efficiency is determined by both codon bias and folding energy. Proc. Natl. Acad. Sci. U. S. A. *107*, 3645–3650.
- Tuller, T., Veksler-Lublinsky, I., Gazit, N., Kupiec, M., Ruppin, E., and Ziv-Ukelson, M. (2011). Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol *12*, R110.
- Turkina, M. V., Klang Årstrand, H., Vener, A. V., Årstrand, H., and Vener, A. V. (2011). Differential phosphorylation of ribosomal proteins in Arabidopsis thaliana plants during day and night. PLoS One *6*, e29307.
- Tzani, I., Ivanov, I.P., Andreev, D.E., Dmitriev, R.I., Dean, K.A., Baranov, P. V, Atkins, J.F., and Loughran, G. (2016). Systematic analysis of the PTEN 5′ leader identifies a major AUU initiated proteoform. Open Biol. *6*, 150203.
- Uchil, P.D., and Satchidanandam, V. (2003). Architecture of the flaviviral replication complex: Protease, nuclease, and detergents reveal encasement within double-layered membrane compartments. J. Biol. Chem. *278*, 24388–24398.
- Ude, S., Lassak, J., Starosta, A.L., Kraxenberger, T., Wilson, D.N., and Jung, K. (2013). Translation elongation factor EF-P alleviates ribosome stalling at polyproline stretches. Science *339*, 82–85.
- Umareddy, I., Chao, A., Sampath, A., Gu, F., and Vasudevan, S.G. (2006). Dengue virus NS4B interacts with NS3 and dissociates it from single-stranded RNA. J. Gen. Virol. *87*, 2605–2614.
- Vanderperre, B., Lucier, J.F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., Wisztorski, M., Salzet, M., Boisvert, F.M., and Roucou, X. (2013). Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. PLoS One *8*.
- Vasquez, J.J., Hon, C.C., Vanselow, J.T., Schlosser, A., and Siegel, T.N. (2014). Comparative ribosome profiling reveals extensive translational complexity in different Trypanosoma brucei life cycle stages. Nucleic Acids Res. *42*, 3623–3637.

- Vattem, K.M., and Wek, R.C. (2004). Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. Proc. Natl. Acad. Sci. U. S. A. *101*, 11269–11274.
- Vaughn, J.N., Ellingson, S.R., Mignone, F., and Arnim, A. von (2012). Known and novel post-transcriptional regulatory sequences are conserved across plant families. RNA *18*, 368–384.
- Versteeg, G.A., van de Nes, P.S., Bredenbeek, P.J., and Spaan, W.J.M. (2007). The coronavirus spike protein induces endoplasmic reticulum stress and upregulation of intracellular chemokine mRNA concentrations. J. Virol. *81*, 10981–10990.
- Viktorovskaya, O. V., Greco, T.M., Cristea, I.M., and Thompson, S.R. (2016). Identification of RNA Binding Proteins Associated with Dengue Virus RNA in Infected Cells Reveals Temporally Distinct Host Factor Requirements. PLoS Negl. Trop. Dis. *10*, e0004921.
- Villordo, S.M., and Gamarnik, A. V (2013). Differential RNA Sequence Requirement for Dengue Virus Replication in Mosquito and Mammalian Cells. J. Virol. *87*, 9365–9372.
- Villordo, S.M., and Gamarnik, A. V. (2009). Genome cyclization as strategy for flavivirus RNA replication. Virus Res. *139*, 230–239.
- Villordo, S.M., Alvarez, D.E., and Gamarnik, A. V (2010). A balance between circular and linear forms of the dengue virus genome is crucial for viral replication. RNA *16*, 2325–2335.
- Villordo, S.M., Filomatori, C. V., Sanchez-Vargas, I., Blair, C.D., and Gamarnik, A. V. (2015). Dengue virus RNA structure specialization facilitates host adaptation. PLoS Pathog. *11*, e1004604.
- Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. *13*, 227–232.
- Vogel, C., Abreu, R. de S. e. S., Ko, D., Le, S.-Y.S.-. Y., Shapiro, B.A., Burns, S.C., de Sousa Abreu, R., Ko, D., Le, S.-Y.S.-. Y., Shapiro, B.A., et al. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol Syst Biol *6*, 400.
- Wallace, E.W.J., Airoldi, E.M., and Drummond, D.A. (2013). Estimating selection on synonymous codon usage from noisy experimental data. Mol. Biol. Evol. *30*, 1438–1453.
- Walsh, D., and Mohr, I. (2011). Viral subversion of the host protein synthesis machinery. Nat. Rev. Microbiol. *9*, 860–875.
- Walsh, D., Mathews, M.B., and Mohr, I. (2013). Tinkering with translation: protein synthesis in virus-infected cells. (Cold Spring Harbor Laboratory Press).
- Wan, J., and Qian, S.-B. (2014). TISdb: a database for alternative translation initiation in mammalian cells. Nucleic Acids Res. *42*, D845–D850.
- Wan, Y., Qu, K., Zhang, Q.C., Flynn, R.A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R.C., Snyder, M.P., Segal, E., et al. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. Nature *505*, 706–709.
- Wang, C., Han, B., Zhou, R., and Zhuang, X. (2016). Real-Time Imaging of Translation on Single mRNA Transcripts in Live Cells. Cell *165*, 990–1001.
- Wang, X., Lu, Z., Gomez, A., Hon, G.C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., et al. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. Nature *505*, 117–120.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

- Watatani, Y., Ichikawa, K., Nakanishi, N., Fujimoto, M., Takeda, H., Kimura, N., Hirose, H., Takahashi, S., and Takahashi, Y. (2008). Stress-induced translation of ATF5 mRNA is regulated by the 5'-untranslated region. J. Biol. Chem. *283*, 2543–2553.
- Weinberg, D.E., Shah, P., Eichhorn, S.W., Hussmann, J.A., Plotkin, J.B., and Bartel, D.P. (2016). Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. Cell Rep *14*, 1787–1799.
- Weiss, R.B., Huang, W.M., and Dunn, D.M. (1990). A nascent peptide is required for ribosomal bypass of the coding gap in bacteriophage T4 gene 60. Cell *62*, 117–126.
- Weisser, M., Voigts-Hoffmann, F., Rabl, J., Leibundgut, M., and Ban, N. (2013). The crystal structure of the eukaryotic 40S ribosomal subunit in complex with eIF1 and eIF1A. Nat. Struct. Mol. Biol. *20*, 1015–1017.
- Wells, S.E., Hillner, P.E., Vale, R.D., and Sachs, A.B. (2016). Circularization of mRNA by Eukaryotic Translation Initiation Factors. Mol. Cell *2*, 135–140.
- Welsch, S., Miller, S., Romero-Brey, I., Merz, A., Bleck, C.K.E., Walther, P., Fuller, S.D., Antony, C., Krijnse-Locker, J., and Bartenschlager, R. (2009). Composition and Three-Dimensional Architecture of the Dengue Virus Replication and Assembly Sites. Cell Host Microbe *5*, 365–375.
- Wethmar, K. (2014). The regulatory potential of upstream open reading frames in eukaryotic gene expression. Wiley Interdiscip. Rev. RNA *5*, 765–768.
- Whitehead, S.S., Blaney, J.E., Durbin, A.P., and Murphy, B.R. (2007). Prospects for a dengue virus vaccine. Nat Rev Microbiol *5*, 518–528.
- Wiborg, O., Pedersen, M.S., Wind, A., Berglund, L.E., Marcker, K.A., and Vuust, J. (1985). The human ubiquitin multigene family: some genes contain multiple directly repeated ubiquitin coding sequences. EMBO J. *4*, 755–759.
- Wiese, A. (2004). A Conserved Upstream Open Reading Frame Mediates Sucrose-Induced Repression of Translation. Plant Cell Online *16*, 1717–1729.
- Williams, C.C., Jan, C.H., and Weissman, J.S. (2014). Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling. Science. *346*, 748–751.
- Wills, N.M., Moore, B., Hammer, A., Gesteland, R.F., and Atkins, J.F. (2006). A functional -1 ribosomal frameshift signal in the human paraneoplastic Ma3 gene. J. Biol. Chem. *281*, 7082–7088.
- Wilson, D.N., and Cate, J.H.D. (2012). The structure and function of the eukaryotic ribosome. Cold Spring Harb. Perspect. Biol. *4*, 5.
- Winfield, M.O., Lu, C., Wilson, I.D., Coghill, J. a, and Edwards, K.J. (2009). Cold- and light-induced changes in the transcriptome of wheat leading to phase transition from vegetative to reproductive growth. BMC Plant Biol. *9*, 55.
- Winfield, M.O., Lu, C., Wilson, I.D., Coghill, J.A., and Edwards, K.J. (2010). Plant responses to cold: transcriptome analysis of wheat. Plant Biotechnol. J. *8*, 749–771.
- Wohlgemuth, I., Pohl, C., Mittelstaet, J., Konevega, A.L., and Rodnina, M. V (2011). Evolutionary optimization of speed and accuracy of decoding on the ribosome. Philos. Trans. R. Soc. B Biol. Sci. *366*, 2979–2986.
- Wolin, S.L., and Walter, P. (1988). Ribosome pausing and stacking during translation of a eukaryotic mRNA. EMBO J. *7*, 3559–3569.
- Woo, P.C.Y., Lau, S.K.P., Huang, Y., and Yuen, K.-Y. (2009). Coronavirus diversity, phylogeny and interspecies jumping. Exp. Biol. Med. *234*, 1117–1127.

- Woolstenhulme, C.J., Parajuli, S., Healey, D.W., Valverde, D.P., Petersen, E.N., Starosta, A.L., Guydosh, N.R., Johnson, W.E., Wilson, D.N., and Buskirk, A.R. (2013). Nascent peptides that block protein synthesis in bacteria. Proc. Natl. Acad. Sci. *110*, E878–E887.
- Woolstenhulme, C.J., Guydosh, N.R., Green, R., and Buskirk, A.R. (2015). High-Precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. Cell Rep. *11*, 13–21.
- Wu, B., Eliscovich, C., Yoon, Y.J., and Singer, R.H. (2016). Translation dynamics of single mRNAs in live cells and neurons. Science. aaf1084.
- Xiao, Z., Zou, Q., Liu, Y., and Yang, X. (2016). Genome-wide assessment of differential translations with ribosome profiling data. Nat Commun *7*, 11194.
- Xie, S.-Q., Nie, P., Wang, Y., Wang, H., Li, H., Yang, Z., Liu, Y., Ren, J., and Xie, Z. (2016). RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling. Nucleic Acids Res. *44*, D254-8.
- Xu, J., and Chua, N.-H. (2009). Arabidopsis decapping 5 is required for mRNA decapping, P-body formation, and translational repression during postembryonic development. Plant Cell *21*, 3270–3279.
- Xu, J., and Chua, N.H. (2011). Processing bodies and plant development. Curr. Opin. Plant Biol. *14*, 88–93.
- Xu, G. lian, Chen, J., Yang, F., Li, G. qing, Zheng, L. xin, and Wu, Y. zhang (2014). C5a/C5aR pathway is essential for the pathogenesis of murine viral fulminant hepatitis by way of potentiating Fgl2/fibroleukin expression. Hepatology *60*, 114–124.
- Xu, H., Wang, P., Fu, Y., Zheng, Y., Tang, Q., Si, L., You, J., Zhang, Z., Zhu, Y., Zhou, L., et al. (2010). Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. Cell Res. *20*, 445–457.
- Yamada, Y., Liu, X.B., Fang, S.G., Tay, F.P.L., and Liu, D.X. (2009). Acquisition of cell-cell fusion activity by amino acid substitutions in spike protein determines the infectivity of a coronavirus in cultured cells. PLoS One *4*, e6130.
- Yamashita, M., and Emerman, M. (2004). Capsid Is a Dominant Determinant of Retrovirus Infectivity in Nondividing Cells. J. Virol. *78*, 5670–5678.
- Yan, S., Wen, J. Der, Bustamante, C., and Tinoco, I. (2015). Ribosome Excursions during mRNA Translocation Mediate Broad Branching of Frameshift Pathways. Cell *160*, 870–881.
- Yan, X., Hoek, T.A., Vale, R.D., and Tanenbaum, M.E. (2016). Dynamics of Translation of Single mRNA Molecules In Vivo. Cell *165*, 976–989.
- Yang, D., and Leibowitz, J.L. (2015). The structure and functions of coronavirus genomic 3' and 5' ends. Virus Res. *206*, 120–133.
- Yang, C.C., Hsieh, Y.C., Lee, S.J., Wu, S.H., Liao, C.L., Tsao, C.H., Chao, Y.S., Chern, J.H., Wu, C.P., and Yueh, A. (2011). Novel dengue virus-specific NS2B/NS3 protease inhibitor, BP2109, discovered by a high-throughput screening assay. Antimicrob. Agents Chemother. *55*, 229–238.
- Yang, J.-. R., Chen, X., and Zhang, J. (2014). Codon-by-Codon Modulation of Translational Speed and Accuracy Via mRNA Folding. PLoS Biol *12*, e1001910.
- Yang, X.-Y.Y., He, K., Du, G., Wu, X., Yu, G., Pan, Y., Zhang, G., Sun, X., and He, Q.-Y.Y. (2016). Integrated translatomics with proteomics to identify novel iron-transporting proteins in Streptococcus pneumoniae. Front. Microbiol. *7*, 78.

- Yang, Z., Cao, S., Martens, C.A., Porcella, S.F., Xie, Z., Ma, M., Shen, B., and Moss, B. (2015). Deciphering Poxvirus Gene Expression by RNA Sequencing and Ribosome Profiling. J. Virol. *89*, 6874–6886.
- Yángüez, E., Castro-Sanz, A.B., Fernández-Bautista, N., Oliveros, J.C., and Castellano, M.M. (2013). Analysis of genome-wide changes in the translatome of Arabidopsis seedlings subjected to heat stress. PLoS One *8*, e71425.
- Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. Nucleic Acids Res. *44*, D710–D716.
- Ye, Y., Hauns, K., Langland, J.O., Jacobs, B.L., and Hogue, B.G. (2007). Mouse hepatitis coronavirus A59 nucleocapsid protein is a type I interferon antagonist. J. Virol. *81*, 2554–2563.
- Yeung, Y.S., Yip, C.W., Hon, C.C., Chow, K.Y.C., Ma, I.C.M., Zeng, F., and Leung, F.C.C. (2008). Transcriptional profiling of Vero E6 cells over-expressing SARS-CoV S2 subunit: Insights on viral regulation of apoptosis and proliferation. Virology *371*, 32–43.
- Yi, Z., Lin, W.W., Stunz, L.L., and Bishop, G.A. (2014). Roles for TNF-receptor associated factor 3 (TRAF3) in lymphocyte functions. Cytokine Growth Factor Rev. *25*, 147–156.
- Yocupicio-Monroy, R.M.E., Medina, F., Reyes-del Valle, J., and del Angel, R.M. (2003). Cellular proteins from human monocytes bind to dengue 4 virus minus-strand 3' untranslated region RNA. J. Virol. *77*, 3067–3076.
- Yoshida, H., Matsui, T., Yamamoto, A., Okada, T., and Mori, K. (2001). XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. Cell *107*, 881–891.
- Young, D.J., Guydosh, N.R., Zhang, F., Hinnebusch, A.G., and Green, R. (2015). Rli1/ABCE1 Recycles Terminating Ribosomes and Controls Translation Reinitiation in 3'UTRs In Vivo. Cell *162*, 872–884.
- Yu, L., Nomaguchi, M., Padmanabhan, R., and Markoff, L. (2008). Specific requirements for elements of the 5' and 3' terminal regions in flavivirus RNA synthesis and viral replication. Virology *374*, 170–185.
- Yusupova, G., and Yusupov, M. (2014). High-Resolution Structure of the Eukaryotic 80S Ribosome. Annu. Rev. Biochem. *83*, 467–486.
- Yusupova, G.Z., Yusupov, M.M., Cate, J.H.D., and Noller, H.F. (2001). The path of messenger RNA through the ribosome. Cell *106*, 233–241.
- Zangi, L., Lui, K.O.K., Gise, A. Von, Ma, Q., Ebina, W., von Gise, A., Ma, Q., Ebina, W., Ptaszek, L.M., Später, D., et al. (2013). Modified mRNA directs the fate of heart progenitor cells and induces vascular regeneration after myocardial infarction. Nat. Biotechnol. *31*, 898–907.
- Zhang, J., Pan, X., Yan, K., Sun, S., Gao, N., and Sui, S.-F.F. (2015). Mechanisms of ribosome stalling by SecM at multiple elongation steps. Elife *4*, e09684.
- Zhang, Y., Wulfle, T., and Rospert, S. (2013). Interaction of nascent chains with the ribosomal tunnel proteins Rpl4, Rpl17, and Rpl39 of saccharomyces cerevisiae. J. Biol. Chem. *288*, 33697–33707.
- Zhang, Z.-H., Wray, N.R., and Zhao, Q. (2016). DEAR-O: Differential Expression Analysis based on RNA-seq data - Online. bioRxiv *2*, 69807.
- Zhao, C., Datta, S., Mandal, P., Xu, S., and Hamilton, T. (2010). Stress-sensitive regulation of IFRD1 mRNA decay is mediated by an upstream open reading frame. J. Biol. Chem. *285*, 8552–8562.

- Zhao, Y., Soh, T.S., Zheng, J., Chan, K.W.K., Phoo, W.W., Lee, C.C., Tay, M.Y.F., Swaminathan, K., Cornvik, T.C., Lim, S.P., et al. (2015). A Crystal Structure of the Dengue Virus NS5 Protein Reveals a Novel Inter-domain Interface Essential for Protein Flexibility and Virus Replication. PLoS Pathog. *11*, 1–27.
- Zheng, W., Chung, L.M., and Zhao, H. (2011). Bias detection and correction in RNA-Sequencing data. BMC Bioinformatics *12*, 290.
- Zhong, Y., Karaletsos, T., Drewe, P., T Sreedharan, V.T., Kuo, D., Singh, K., Wendel, H.-G., and Rätsch, G. (2016). RiboDiff: Detecting Changes of Translation Efficiency from Ribosome Footprints. bioRxiv.
- Zhou, D., Palam, L.R., Jiang, L., Narasimhan, J., Staschke, K.A., and Wek, R.C. (2008). Phosphorylation of eIF2 directs ATF5 translational control in response to diverse stress conditions. J. Biol. Chem. *283*, 7064–7073.
- Zhu, J., Dong, C.H., and Zhu, J.K. (2007). Interplay between cold-responsive gene regulation, metabolism and RNA processing during plant cold acclimation. Curr. Opin. Plant Biol. *10*, 290–295.
- Zorzitto, J., Galligan, C.L., Ueng, J.J.M., and Fish, E.N. (2006). Characterization of the antiviral effects of interferon-alpha against a SARS-like coronoavirus infection in vitro. Cell Res. *16*, 220–229.
- Zoschke, R., Watkins, K.P., and Barkan, A. (2013). A Rapid Ribosome Profiling Method Elucidates Chloroplast Ribosome Behavior in Vivo. Plant Cell *25*, 2265–2275.
- Zúñiga, S., Sola, I., Alonso, S., and Enjuanes, L. (2004). Sequence Motifs Involved in the Regulation of Discontinuous Coronavirus Subgenomic RNA Synthesis. J. Virol. *78*, 980–994.
- Zur, H., and Tuller, T. (2012). Strong association between mRNA folding strength and protein abundance in S. cerevisiae. EMBO Rep. *13*, 272–277.